CONSTRUCTION AND CONVERGENCE STUDY OF SCHEMES PRESERVING THE ELLIPTIC LOCAL MAXIMUM PRINCIPLE

Jerome Droniou, Christophe Le Potier

▶ **To cite this version:**

Jerome Droniou, Christophe Le Potier. CONSTRUCTION AND CONVERGENCE STUDY OF SCHEMES PRESERVING THE ELLIPTIC LOCAL MAXIMUM PRINCIPLE. SIAM Journal on Numerical Analysis, 2011, 49 (2), pp.459-490. 10.1137/090770849 . hal-00808694

HAL Id: hal-00808694
https://hal.science/hal-00808694

Submitted on 6 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SIAM J. NUMER. ANAL.
Vol. 49, No. 2, pp. 459–490
© 2011 Society for Industrial and Applied Mathematics

CONSTRUCTION AND CONVERGENCE STUDY OF SCHEMES PRESERVING THE ELLIPTIC LOCAL MAXIMUM PRINCIPLE*

JÉRÔME DRONIOU† AND CHRISTOPHE LE POTIER‡

Abstract. We present a method to approximate (in any space dimension) diffusion equations with schemes having a specific structure; this structure ensures that the discrete local maximum and minimum principles are respected, and that no spurious oscillations appear in the solutions. When applied in a transient setting on models of concentration equations, it guaranties in particular that the approximate solutions stay between the physical bounds. We make a theoretical study of the constructed schemes, proving under a coercivity assumption that their solutions converge to the solution of the PDE. Several numerical results are also provided; they help us understand how the parameters of the method should be chosen. These results also show the practical efficiency of the method, even when applied to complex models.

Key words. finite volumes, anisotropic heterogeneous diffusion, maximum principle, convergence study, numerical tests

AMS subject classifications. 65N08, 65N12

DOI. 10.1137/090770849

1. Introduction. Let Ω be an open bounded connected polygonal domain of $\mathbb{R}^N$. We consider the following problem:

$$\left\{\begin{array}{ll} \vec{q} = -\mathbf{D}\nabla u & \text{in } \Omega\,, \\ \operatorname{div} \vec{q} = f & \text{in } \Omega\,, \\ u = u_\partial & \text{on } \partial\Omega \end{array}\right. \tag{1}$$

with

(i) f, the source term, belonging to $L^2(\Omega)$;

(ii) u, the concentration of the radioactive element;

(iii) $\mathbf{D}$, the permeability, a symmetric tensor-valued function such that (a) $\mathbf{D}$ is piecewise Lipschitz-continuous on Ω and (b) the set of the eigenvalues of $\mathbf{D}(x)$ is included in $[\lambda_{\min}, \lambda_{\max}]$ (with $\lambda_{\min} > 0$) for all $x \in \Omega$;

(iv) u_∂, the boundary data.

This basic equation (or its transient version) is at the core of complex models of flows in porous media, used, for example, in petroleum engineering or in the framework of nuclear waste disposal. In such situations, it is crucial to have robust approximations of the solution to (1). This robustness is, in particular, measured through the respect of the physical bounds; for instance, in models of two-phase flows in porous media [36], ensuring that the numerically computed concentration stays between 0 and 1 is of utmost importance; this is also the case when coupling transport equations with chemical models.

If the grid used for the discretization of the PDE has specific orthogonality conditions (depending on $\mathbf{D}$), the classical finite volume scheme [22] (in which numerical

*Received by the editors September 11, 2009; accepted for publication (in revised form) November 29, 2010; published electronically March 15, 2011. This work was supported by GDR MOMAS CNRS/PACEN and Project VFSitCom (ANR-08-BLAN-0275-01).

http://www.siam.org/journals/sinum/49-2/77084.html

†Université Montpellier 2, Institut de Mathématiques et de Modélisation de Montpellier, CC 051, Place Eugène Bataillon, 34095 Montpellier cedex 5, France (droniou@math.univ-montp2.fr).

‡CEA-Saclay, DEN, DM2S, SFME, F-91191 Gif-sur-Yvette, France (clepotier@cea.fr).

fluxes are approximated by a two-point finite difference expression) provides a solution that respects these bounds. However, in practical situations such a very specific grid is not available, and might not even be constructible. Several methods have been recently developed to construct schemes for elliptic PDEs on generic meshes: multipoint flux approximations [1, 2, 3], discontinuous Galerkin [17, 37], discrete duality finite volume [18, 8], mimetic finite difference [9, 10, 5], hybrid finite volume [23] (see also the cell-centered variant [24]), mixed finite volume [19, 21] (these last three turn out to be identical [20]). None of these methods, however, ensures that the abovementioned physical bounds are satisfied in every situation, as shown in the FVCA5 benchmark organized in 2008 [25].

Some methods have been specifically designed to discretize (1) by ensuring that the approximate solution satisfies a discrete maximum and/or minimum principle. As it has been proved in [28, 11, 27], no *linear* consistent nine-point control volume scheme constructed on square meshes with a very anisotropic tensor (or on very distorted quadrangular cells with an isotropic tensor) can respect the maximum or minimum principle. One must therefore look for *nonlinear* schemes in order to satisfy these principles.

In [12], a nonlinear correction of the classical linear P_1 finite element is proposed, but heterogeneous anisotropic tensors are not taken into account. In [6], an interesting nonlinear method is proposed for homogeneous isotropic diffusions. Unfortunately, the positivity properties are obtained under restrictive geometric constraints. In [30], a cell-centered finite volume discretization for diffusion operators is proposed, and its robustness and accuracy are shown through comparisons with analytical solutions. This scheme satisfies either the minimum or the maximum principle but not both principles simultaneously; it has been extended in [26, 34, 35, 38] on polygonal meshes and tetrahedrons. However, no scheme developed in these articles satisfies with certainty both the minimum and the maximum principles simultaneously, and no theoretical proof of convergence is given.

We can also cite the recent work [32] on a linear scheme satisfying a maximum principle for anisotropic diffusion operators on distorted grids in dimension 2. Unfortunately, this method is, in general, only of order 1 in space and has a nonstandard stencil.

In [33], a new finite volume method for highly anisotropic diffusion operators is introduced on triangular meshes. This scheme satisfies a discrete version of the classical local maximum (and minimum) principle for elliptic equations without geometrical constraints on the mesh and restrictive conditions on the anisotropy ratio.

Our goal in the present work is to extend this method to very generic grids in any space dimension. Moreover, we improve the precision of the scheme in some cases of discontinuous diffusion tensor. We also prove its convergence toward the solution of (1) as the size of the mesh tends to 0; this theoretical study is not just a mathematical amusement since it leads us to an understanding of how to choose the parameters of the method in order to obtain good approximations of the solution.

The finite volume framework is chosen because it ensures that the approximation satisfies the local conservation of mass, an essential physical property. The maximum and minimum principles are harder to satisfy, and thus more rarely considered in the construction of finite volume schemes; they are, however, also quite important: they ensure not only that the approximate concentration stays within the physical bounds, but also that it does not develop spurious oscillations. The schemes that satisfy these principles are therefore very robust.

The articulation of the paper is as follows. In the rest of this introduction, we define a structure of schemes that ensures that the discrete maximum and minimum principles are respected; this structure also makes sure that the approximate solution does not develop spurious oscillations. In section 2, we present a method to construct schemes satisfying this structure on generic meshes, while having a simple nine-point stencil on most quadrangular grids (even in the presence of strong anisotropy). We start with the simpler case of the isotropic homogeneous tensor $\mathbf{D} = \mathrm{Id}$ before detailing the case of a generic anisotropic heterogeneous tensor. In section 3, the theoretical study of the obtained schemes is developed, from the proof of the existence of solutions to their convergence toward the continuous solution (as the mesh size tends to 0). As this is the case for other schemes based on flux balances and multipoint approximations of the fluxes, the coercivity of the method cannot be theoretically ensured in any situation; the proof of convergence is thus made under a coercivity assumption. The numerical results we present in section 4, however, show that this assumption seems to hold quite well in practice, even for strongly anisotropic and heterogeneous permeabilities. In this section, we also take advantage of findings made during the theoretical study to present adequate choices for the parameters of the method. We discuss in particular the case of discontinuous diffusion tensors; this case is very important in practice, but few papers in the literature on monotone schemes seem to fully take it into account. Finally, we numerically illustrate the efficiency of our method by comparing it with other (linear) schemes in the case of a strongly discontinuous diffusion tensor, and on the more realistic COUPLEX 1 benchmark from [7]. A short conclusion closes the article.

1.1. The local maximum principle structure. The basic assumptions and notation on the discretization of Ω are the following.

DEFINITION 1.1 (admissible mesh). *An admissible mesh of Ω is triplet $\mathcal{D} = (\mathcal{T}, \mathcal{A}, \mathcal{P})$ where*

1. *$\mathcal{T}$ is a finite family of nonempty connected open disjoint subsets of Ω (the* cells *or* control volumes*) such that $\overline{\Omega} = \cup_{K\in\mathcal{T}} \overline{K}$;*

2. *$\mathcal{A}$ is a finite family of subsets of $\overline{\Omega}$ (the* edges*—faces in dimension 3) such that any $a \in \mathcal{A}$ is a nonempty closed subset of a hyperplane of $\mathbb{R}^N$ with positive $(N-1)$-dimensional measure, and such that the intersection of two different edges has zero $(N-1)$-dimensional measure. We also assume that, for all $K \in \mathcal{T}$, there exists a subset $\mathcal{A}_K$ of $\mathcal{A}$ such that $\partial K = \cup_{a\in\mathcal{A}_K} a$ and that any edge is contained either in $\partial\Omega$ or in $\mathcal{A}_K \cap \mathcal{A}_L$ for two distinct control volumes K and L;*

3. *$\mathcal{P} = (X_K)_{K\in\mathcal{T}}$ is a family of points (the* cell centers*—not necessarily the center of gravity of the cells) of Ω such that, for all $K \in \mathcal{T}$, $X_K \in K$.*

Remark 1.2. Notice that, although it is not mandatory for the construction and study of the scheme, the considered meshes are nearly always aligned with the discontinuities of $\mathbf{D}$ (i.e., for all $K \in \mathcal{T}$, $\mathbf{D}$ is continuous on K). The discontinuities of $\mathbf{D}$ are in general due to the geological layers, and the mesh usually follows these layers; hence, the alignment of the mesh with these discontinuities is quite natural.

For all $K \in \mathcal{T}$, $|K|$ is the N-dimensional measure of K. For $a \in \mathcal{A}$, $|a|$ is the $(N-1)$-dimensional measure of a. The edges a contained in $\partial\Omega$ are called boundary edges, with the other edges being interior edges; $\mathcal{A}_{\mathrm{ext}}$ is the set of boundary edges, and $\mathcal{A}_{\mathrm{int}}$ is the set of interior edges. If $a \in \mathcal{A}_K$, the unit normal to a in the outer direction of K is $\vec{n}_{K,a}$. The size of the mesh is $\mathrm{size}(\mathcal{D}) = \sup_{K\in\mathcal{T}} \mathrm{diam}(K)$, where $\mathrm{diam}(K)$ is the diameter of K. The Euclidean distance is denoted by d, and the Euclidean norm of a vector $\vec{v}$ is written $|\vec{v}|$.

A numerical scheme for (1) is an equation on some unknowns $(u_K)_{K\in\mathcal{T}}$, assumed to be approximate values of the solution to this PDE at the cell centers $(X_K)_{K\in\mathcal{T}}$.

DEFINITION 1.3 (LMP structure). *Let $\mathcal{D}$ be an admissible mesh of Ω. We say that a scheme for (1) using the unknowns $u = (u_K)_{K\in\mathcal{T}}$ has the local maximum principle structure (LMP structure for short) if it can be written in the form*

$$\forall K \in \mathcal{T} : \sum_{L\in\mathcal{T}} \tau_{K,L}(u)(u_K - u_L) + \sum_{a\in\mathcal{A}_{\rm ext}} \tau_{K,a}(u)(u_K - u_a) = \int_K f \tag{2}$$

for some functions $\tau_{K,L} : \mathbb{R}^{\mathrm{Card}(\mathcal{T})} \mapsto \mathbb{R}^+$ (for $(K,L) \in \mathcal{T}^2$) and $\tau_{K,a} : \mathbb{R}^{\mathrm{Card}(\mathcal{T})} \mapsto \mathbb{R}^+$ (for $K \in \mathcal{T}$ and $a \in \mathcal{A}_{\rm ext}$) satisfying, for all $v \in \mathbb{R}^{\mathrm{Card}(\mathcal{T})}$,

$$\begin{aligned} \forall (K,L) \in \mathcal{T}^2 \text{ such that } \mathcal{A}_K \cap \mathcal{A}_L \neq \emptyset : \quad & \tau_{K,L}(v) > 0, \\ \forall K \in \mathcal{T},\ \forall a \in \mathcal{A}_K \cap \mathcal{A}_{\rm ext} : \quad & \tau_{K,a}(v) > 0. \end{aligned} \tag{3}$$

In (2), u_a stands for some value of u_∂ on a; we assume that this value is between the minimum and maximum values of u_∂.

This definition is set in order that the classical proofs of the discrete maximum and minimum principles (proofs that are used, for example, in the case of two-point fluxes finite volume methods) can be straightforwardly adapted to schemes having the LMP structure. It is also important to notice that the LMP structure ensures that both principles simultaneously hold; this is not the case for most schemes built on general grids: as pointed out in the introduction, schemes on general grids oftentimes enjoy only one, or none, of these two principles.

PROPOSITION 1.4 (discrete local maximum and minimum principles). *Assume that $u_\partial = 0$. If $f \geq 0$ (resp., $f \leq 0$) and $u = (u_K)_{K\in\mathcal{T}}$ is a solution to a scheme having the LMP structure then $\min_{K\in\mathcal{T}} u_K \geq 0$ (resp., $\max_{K\in\mathcal{T}} u_K \leq 0$).*

Moreover, if this minimum (resp., maximum) is attained in a cell K such that $\tau_{K,a}(\cdot) = 0$ for all $a \in \mathcal{A}_{\rm ext}$, then u is locally constant around K: the values of u in this cell and all the neighboring cells are identical. In particular, if the scheme is such that $\tau_{K,a}(\cdot) = 0$ for all $K \in \mathcal{T}$ and all $a \notin \mathcal{A}_K$ then, unless it is constant, u cannot attain its minimum (resp., maximum) in an interior cell (i.e., a cell K such that $\mathcal{A}_K \cap \mathcal{A}_{\rm ext} = \emptyset$).

Similarly, the proof of the following nonoscillating property is an easy consequence of the LMP structure.

PROPOSITION 1.5 (nonoscillating property). *Let $f = 0$ and $u = (u_K)_{K\in\mathcal{T}}$ be a solution to a scheme having the LMP structure; for $K \in \mathcal{T}$, we define $V(K) = \{L \in \mathcal{T} \mid \tau_{K,L}(u) \neq 0\}$ and $E(K) = \{a \in \mathcal{A}_{\rm ext} \mid \tau_{K,a}(u) \neq 0\}$. Then, for any cell K, we have $\min(\min_{J\in V(K)} u_J, \min_{a\in E(K)} u_a) \leq u_K \leq \max(\max_{J\in V(K)} u_J, \max_{a\in E(K)} u_a)$.*

Once a scheme for the stationary equation (1) is available, it is straightforward to build a scheme for the corresponding parabolic equation

$$\begin{cases} \partial_t u = \mathrm{div}(\mathbf{D}\nabla u) & \text{in }]0,\infty[\times\Omega, \\ u = u_\partial & \text{on }]0,\infty[\times\partial\Omega, \\ u = u_0 & \text{on } \{0\}\times\Omega. \end{cases} \tag{4}$$

Assume that a scheme for (1) is written $S_\mathcal{D}(u) = (\int_K f)_{K\in\mathcal{T}}$ with $S_\mathcal{D} : \mathbb{R}^{\mathrm{Card}(\mathcal{T})} \to \mathbb{R}^{\mathrm{Card}(\mathcal{T})}$ (in the case of an LMP structure, $S_\mathcal{D}(u)_K$ is the left-hand side of (2)); then, using a time-implicit discretization, a scheme for (4) is given by

$$\begin{aligned} &\forall n \geq 0,\ \forall K \in \mathcal{T} :\ |K| u_K^{n+1} + \delta t S_\mathcal{D}(u^{n+1})_K = |K| u_K^n, \\ &\forall K \in \mathcal{T} :\ u_K^0 = \frac{1}{|K|}\int_K u_0. \end{aligned} \tag{5}$$

Another easy property of schemes having the LMP structure is the following.

PROPOSITION 1.6 (preservation of the initial bounds). *Assume that* $0 \le u_0 \le 1$ *and that* $0 \le u_\partial \le 1$. *If* $S_\mathcal{D}$ *defines a scheme having the LMP structure and if* $u = (u_K^n)_{K\in\mathcal{T},\, n\ge 0}$ *is a solution to* (5), *then for all* $n \ge 0$ *and all* $K \in \mathcal{T}$ *we have* $0 \le u_K^n \le 1$.

2. Presentation of the method. The construction of schemes having the LMP structure demands some local geometrical reasoning and is a little bit easier to present in the isotropic homogeneous case $\mathbf{D} = \mathrm{Id}$. We thus first handle this case before turning to the more general diffusion tensor.

In the following sections, we present a family of schemes, each one corresponding to specific choices of parameters that appear during the construction. We discuss these choices in section 4, using the theoretical study of section 3 as a guide to select "proper" parameters.

2.1. Isotropic homogeneous case: $\mathbf{D} = \mathbf{Id}$. We add the following assumption on the mesh, the statement of which is easily understood by looking at Figure 1.

Assumption 2.1.

1. $\mathcal{D}$ is an admissible mesh.

2. For all boundary edges $b \in \mathcal{A}_{\mathrm{ext}}$, denoting by $T(b)$ the control volume such that $b \in \mathcal{A}_{T(b)}$, we assume that the half-line starting at $X_{T(b)}$ in the direction $\vec{n}_{T(b),b}$ intersects b at some point denoted by X_b.

3. For all $a \in \mathcal{A}_{\mathrm{int}}$, we denote by $T_1(a)$ and $T_2(a)$ the control volumes on either side of a, and we take two points $M_{1,a}$ and $M_{2,a}$ inside the convex hull of the cell centers and boundary points $((X_K)_{K\in\mathcal{T}}, (X_b)_{b\in\mathcal{A}_{\mathrm{ext}}})$ such that

(i) $M_{2,a}$ belongs to the (open) half-line starting at $X_{T_1(a)}$ and with direction $\vec{n}_{T_1(a),a}$;

(ii) $M_{1,a}$ belongs to the (open) half-line starting at $X_{T_2(a)}$ and with direction $\vec{n}_{T_2(a),a}$.

We let $\mathcal{M}$ be the set of the chosen points $(M_{i,a})_{a\in\mathcal{A}_{\mathrm{int}},\, i=1,2}$.

Remark 2.2. Note that Assumption 2.1 does not require that the straight lines mentioned in items (3.i) and (3.ii) intersect the interior edge a. This intersection is mandatory only for *boundary* edges in item (2); one can, for example, see in Figure 1 that the straight line starting at $X_{T(b)}$ and directed by $\vec{n}_{T(b),c}$ does not intersect the edge c. See also Remark 2.7.

We then describe in three steps a method for constructing a scheme having the LMP structure. This scheme is written using the unknowns $(u_K)_{K\in\mathcal{T}}$ playing the role of approximate values of the solution at the cell centers $(X_K)_{K\in\mathcal{T}}$.

In the following, we denote $u_b = u_\partial(X_b)$.

Step 1 *(interpolation of additional approximate values of the solution).* The points $M_{i,a}$ given by Assumption 2.1 can be written as convex combinations of cell centers and boundary points. This allows us to define, by interpolation of $(u_K)_{K\in\mathcal{T}}$ and $(u_b)_{b\in\mathcal{A}_{\mathrm{ext}}}$, some approximate values $u_{M_{i,a}}$ of the solution at $M_{i,a}$.

In the convex combination used to write $M_{i,a}$, the coefficient of $X_{T_i(a)}$ plays a particular role; we therefore give it a special name, say $\alpha_{i,a}$. We thus have $M_{i,a} = \alpha_{i,a}X_{T_i(a)} + \sum_{j=1}^{J_{i,a}} \lambda_{i,a}(j)X_{i,a}(j)$ (with $\alpha_{i,a} \ge 0$, $\lambda_{i,a}(j) \ge 0$, and $\alpha_{i,a} + \sum_{j=1}^{J_{i,a}} \lambda_{i,a}(j) = 1$, and, for each j, $X_{i,a}(j)$ is either X_K for some $K \in \mathcal{T}$ or X_b for some $b \in \mathcal{A}_{\mathrm{ext}}$). It is then natural to define

$$u_{M_{i,a}} = \alpha_{i,a}u_{T_i(a)} + \sum_{j=1}^{J_{i,a}} \lambda_{i,a}(j)u_{X_{i,a}(j)}, \tag{6}$$

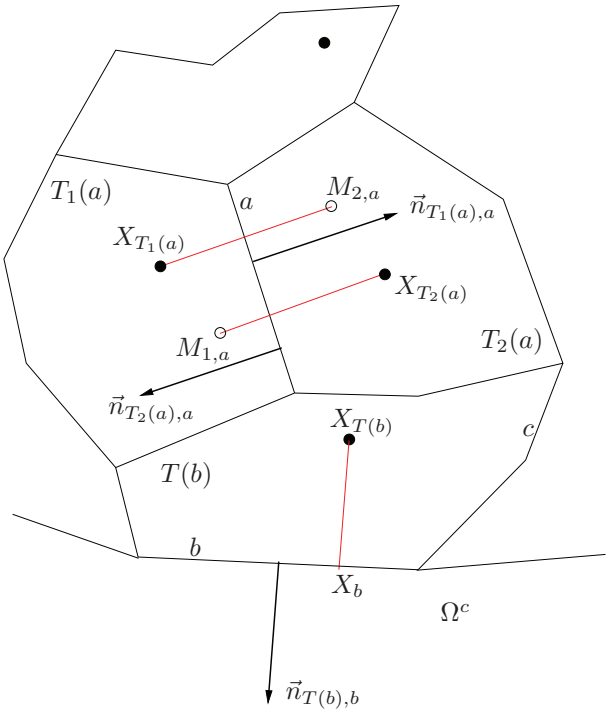

FIG. 1. *Illustration of Assumption* 2.1 *(choice of* $M_{i,a}$ *in the homogeneous case).*

with the following obvious notation: $u_{X_{i,a}(j)} = u_K$ if $X_{i,a}(j) = X_K$ with $K \in \mathcal{T}$ and $u_{X_{i,a}(j)} = u_b$ if $X_{i,a}(j) = X_b$ with $b \in \mathcal{A}_{\rm ext}$. We assume that $\alpha_{i,a} > 0$, which is usually not a restriction.

Remark 2.3. It is always possible to choose the points $X_{i,a}(j)$ such that the distance between these points and $M_{i,a}$ is of order $\mathrm{diam}(T_i(a))$. In fact, the points $X_{i,a}(j)$ lay, in general, within a few cells of $T_i(a)$.

Step 2 *(definition of the approximate value of the flux* $\vec{q}\cdot\vec{n}_{K,a}$*).* This is of course the key step in the construction of the scheme. In the following, the dependence of a quantity Q on $u = (u_K)_{K\in\mathcal{T}}$ is explicitly indicated by $Q(u)$; quantities that are not indicated as such depend only on the mesh.

Let $K \in \mathcal{T}$, and let $a \in \mathcal{A}_K$ be a boundary edge. Taking into account Assumption 2.1 and the vanishing boundary value in (1), a consistent approximation of $\vec{q}\cdot\vec{n}_{K,a}$ is

$$F_{K,a}(u) = \frac{u_K - u_a}{d(X_K, X_a)}. \tag{7}$$

Let us now assume that $a \in \mathcal{A}_K$ is an interior edge. Thanks to Assumption 2.1, using $u_{T_1(a)}$, $u_{T_2(a)}$, and the values $u_{M_{i,a}}$ previously defined, we have two consistent ways, $F_{1,a}(u)$ and $F_{2,a}(u)$, to approximate the flux of $\vec{q} = -\nabla u$ through a (one outside $T_1(a)$, and the other outside $T_2(a)$):

$$F_{1,a}(u) = \frac{u_{T_1(a)} - u_{M_{2,a}}}{d(X_{T_1(a)}, M_{2,a})} \quad \text{and} \quad F_{2,a}(u) = \frac{u_{T_2(a)} - u_{M_{1,a}}}{d(X_{T_2(a)}, M_{1,a})}.$$

Owing to (6), this means that

$$F_{1,a}(u) = \frac{\alpha_{2,a}(u_{T_1(a)} - u_{T_2(a)}) + \sum_{j=1}^{J_{2,a}} \lambda_{2,a}(j)(u_{T_1(a)} - u_{X_{2,a}(j)})}{d(X_{T_1(a)}, M_{2,a})} \tag{8}$$

and

$$F_{2,a}(u) = \frac{\alpha_{1,a}(u_{T_2(a)} - u_{T_1(a)}) + \sum_{j=1}^{J_{1,a}} \lambda_{1,a}(j)(u_{T_2(a)} - u_{X_{1,a}(j)})}{d(X_{T_2(a)}, M_{1,a})}. \tag{9}$$

The approximation $F_{K,a}(u)$ chosen for $\vec{q} \cdot \vec{n}_{K,a}$ is a well-chosen combination of these two fluxes. As one might expect, the final finite volume scheme consists of writing the flux balance

$$\forall K \in \mathcal{T} : \sum_{a \in \mathcal{A}_K} |a| F_{K,a}(u) = \int_K f. \tag{10}$$

In order to ensure that this scheme has the LMP structure, we intend on constructing $F_{K,a}(u)$ so that it can be written in the form

$$F_{K,a}(u) = \sum_{L \in \mathcal{T}} \nu_{K,L,a}(u)(u_K - u_L), \tag{11}$$

where $\nu_{K,L,a}(u) \geq 0$ for all L and $\nu_{K,L,a}(u) > 0$ whenever L and K are neighboring cells.[1] If we manage to construct fluxes satisfying (11), equations (7) and (10) clearly show that the resulting scheme has the LMP structure.

Let us first assume that $K = T_1(a)$. We now describe how to choose $\gamma_{1,a}(u)$ and $\gamma_{2,a}(u)$, the coefficients of a convex combination $F_{K,a}(u) = \gamma_{1,a}(u)F_{1,a}(u) + \gamma_{2,a}(u)(-F_{2,a}(u))$ which ensure that (11) holds (recall that $F_{2,a}(u)$ is the flux outside $T_2(a)$, so that $-F_{2,a}(u)$ is the flux outside $T_1(a) = K$ in our current assumption). Fixing $0 < \beta_{1,a} \leq \frac{2\alpha_{1,a}}{d(X_{T_2(a)}, M_{1,a})}$ and $0 < \beta_{2,a} \leq \frac{2\alpha_{2,a}}{d(X_{T_1(a)}, M_{2,a})}$, we notice from (8) and (9) that

$$F_{1,a}(u) = \beta_{2,a}(u_K - u_{T_2(a)}) + G_{1,a}(u) \quad \text{and} \quad -F_{2,a}(u) = \beta_{1,a}(u_K - u_{T_2(a)}) - G_{2,a}(u),$$

where, denoting by "$\star$" some generic nonnegative coefficients,

$$G_{1,a}(u) = \sum_{L \in \mathcal{T}} \star(u_K - u_L) + \left(\frac{\alpha_{2,a}}{d(X_{T_1(a)}, M_{2,a})} - \beta_{2,a} \right) (u_K - u_{T_2(a)}),$$

$$G_{2,a}(u) = \sum_{L \in \mathcal{T}} \star(u_{T_2(a)} - u_L) + \left(\frac{\alpha_{1,a}}{d(X_{T_2(a)}, M_{1,a})} - \beta_{1,a} \right) (u_{T_2(a)} - u_K).$$

Of all the terms $\beta_{2,a}(u_K - u_{T_2(a)})$, $G_{1,a}(u)$, $\beta_{1,a}(u_K - u_{T_2(a)})$, and $-G_{2,a}(u)$ appearing in $F_{1,a}(u)$ and $-F_{2,a}(u)$, only $-G_{2,a}(u)$ and the second part of $G_{1,a}(u)$ have a "bad" form (not similar to (11)). The coefficients $\gamma_{1,a}(u)$ and $\gamma_{2,a}(u)$ are thus chosen in order to give the necessary weight to $F_{1,a}(u)$ (and thus to $G_{1,a}(u)$) and to cancel out this possible bad term $G_{2,a}(u)$: the larger $G_{2,a}(u)$ is, the more we put weight on $F_{1,a}(u)$. We take

$$\gamma_{1,a}(u) = \frac{|G_{2,a}(u)|}{|G_{1,a}(u)| + |G_{2,a}(u)|} \quad \text{and} \quad \gamma_{2,a}(u) = \frac{|G_{1,a}(u)|}{|G_{1,a}(u)| + |G_{2,a}(u)|}$$

[1] In general, $\nu_{K,L,a}(u) = 0$ if L and K are far apart.

(if $|G_{1,a}(u)| + |G_{2,a}(u)| = 0$, we let $\gamma_{1,a}(u) = \gamma_{2,a}(u) = \frac{1}{2}$). The total flux $F_{K,a}(u) = \gamma_{1,a}(u)F_{1,a}(u) + \gamma_{2,a}(u)(-F_{2,a}(u))$ can then be written

$$F_{K,a}(u) = (\gamma_{1,a}(u)\beta_{2,a} + \gamma_{2,a}(u)\beta_{1,a})(u_K - u_{T_2(a)}) + \frac{|G_{2,a}(u)|G_{1,a}(u) - |G_{1,a}(u)|G_{2,a}(u)}{|G_{1,a}(u)| + |G_{2,a}(u)|}.$$

This expression shows that (11) holds: indeed, if $G_{2,a}(u)$ and $G_{1,a}(u)$ have the same sign, then $|G_{2,a}(u)|G_{1,a}(u) - |G_{1,a}(u)|G_{2,a}(u) = 0$ and

$$F_{K,a}(u) = (\gamma_{1,a}(u)\beta_{2,a} + \gamma_{2,a}(u)\beta_{1,a})(u_K - u_{T_2(a)})$$

and, if $G_{2,a}(u)G_{1,a}(u) < 0$, then we have $|G_{2,a}(u)|G_{1,a}(u) - |G_{1,a}(u)|G_{2,a}(u) = 2|G_{2,a}(u)|G_{1,a}(u)$, and thus

$$\begin{aligned} F_{K,a}(u) &= (\gamma_{1,a}(u)\beta_{2,a} + \gamma_{2,a}(u)\beta_{1,a})(u_K - u_{T_2(a)}) \\ &\quad + \frac{2|G_{2,a}(u)|}{|G_{1,a}(u)| + |G_{2,a}(u)|}\left(\frac{\alpha_{2,a}}{d(X_{T_1(a)}, M_{2,a})} - \beta_{2,a}\right)(u_K - u_{T_2(a)}) \\ &\quad + \sum_{L\in\mathcal{T}} \star(u_K - u_L) \\ &= \left(\gamma_{1,a}(u)\left(\frac{2\alpha_{2,a}}{d(X_{T_1(a)}, M_{2,a})} - \beta_{2,a}\right) + \gamma_{2,a}(u)\beta_{1,a}\right)(u_K - u_{T_2(a)}) \\ &\quad + \sum_{L\in\mathcal{T}} \star(u_K - u_L), \end{aligned}$$

which shows that $F_{K,a}(u)$ satisfies (11) since $0 < \beta_{2,a} \le \frac{2\alpha_{2,a}}{d(X_{T_1(a)}, M_{2,a})}$.

If $K = T_2(a)$, we just define, with the same $\beta_{i,a}$ and $\gamma_{i,a}(u)$ as above (not depending on K), $F_{K,a}(u) = -\gamma_{1,a}(u)F_{1,a}(u) + \gamma_{2,a}(u)F_{2,a}(u)$. In summary, the final approximation of $\vec{q}\cdot\vec{n}_{K,a}$ is

$$F_{K,a}(u) = \gamma_{1,a}(u)\epsilon_{1,K}F_{1,a}(u) + \gamma_{2,a}(u)\epsilon_{2,K}F_{2,a}(u), \tag{12}$$

where $\epsilon_{i,K} = +1$ if $K = T_i(a)$ and $\epsilon_{i,K} = -1$ otherwise. These choices of signs clearly ensure (since $\gamma_{i,a}(u)$ and $F_{i,a}(u)$ depend only on a, not on the cell K) that these fluxes are conservative: if K and L are the control volumes on either side of a, then

$$F_{K,a}(u) + F_{L,a}(u) = 0. \tag{13}$$

Remark 2.4. The parameters $(\beta_{i,a})_{i=1,2}$ can be understood as the coefficients of "two-point flux pieces" in the fluxes $(F_{i,a})_{i=1,2}$; in a sense, they represent the "best part" of these fluxes since, whatever the case $K = T_1(a)$ or $K = T_2(a)$, they always give a good contribution in $F_{K,a}$ with respect to (11). Moreover, in situations where the grid is admissible in the sense of [22], the expected flux $F_{K,a}$ should be a two-point flux (such a flux is the simplest consistent flux which satisfies (11)). However, in general, two-point fluxes are clearly not sufficient to construct consistent approximations; this is why we must take into account the rest of the fluxes, i.e., the terms $(G_{i,a})_{i=1,2}$, and give some importance to these terms in the construction of the complete flux $F_{K,a}$.

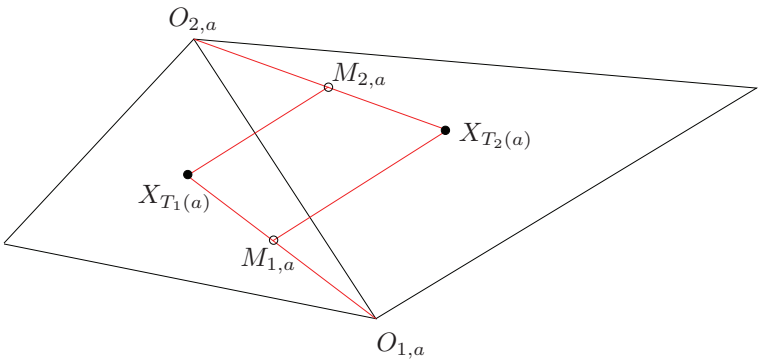

FIG. 2. *Choice of $M_{i,a}$ in the homogeneous case with a triangular mesh.*

Step 3 *(the scheme).* As mentioned above, the scheme for (1) is obtained by writing the flux balance: find $u = (u_K)_{K\in\mathcal{T}}$ such that

$$\forall K \in \mathcal{T} : \sum_{a\in\mathcal{A}_K} |a| F_{K,a}(u) = \int_K f. \tag{14}$$

Remark 2.5 (case of triangular cells in dimension $d = 2$). As described in [33], possible choices for triangular meshes are as follows: the cell centers X_K are the angle bisectors of the triangles, and the points $M_{i,a}$ are chosen on the segments of lines going from these cell centers to the vertices of the edge a (see Figure 2).

Integrating $\operatorname{div} \nabla x = 0$ and $\operatorname{div} \nabla y = 0$ on a disk around the vertex $O_{i,a}$ allows us to express this vertex as a convex combination of the cell centers around it; this gives a similar convex combination for $M_{i,a}$ since

$$M_{i,a} = \mu_{i,a} X_{T_i(a)} + (1 - \mu_{i,a}) O_{i,a}. \tag{15}$$

Finally, we choose

$$\beta_{1,a} = \beta_{2,a} = \min\left(\frac{\mu_{1,a}}{d(X_{T_2(a)}, M_{1,a})}; \frac{\mu_{2,a}}{d(X_{T_1(a)}, M_{2,a})} \right).$$

It can easily be checked (from (15) and writing $O_{i,a}$ as a convex combination of the cell centers around it) that this value is smaller than both $\frac{2\alpha_{1,a}}{d(X_{T_2(a)}, M_{1,a})}$ and $\frac{2\alpha_{2,a}}{d(X_{T_1(a)}, M_{2,a})}$, with $\alpha_{i,a}$ being the coefficient of $X_{T_i(a)}$ in a convex combination giving $M_{i,a}$.

2.2. Anisotropic heterogeneous case. As in the isotropic homogeneous case, we have to introduce an assumption; this is basically the same as Assumption 2.1, except that the directions of the half-lines now follow $\mathbf{D}\vec{n}$ (with $\mathbf{D}$ varying inside each cell). We let $\mathbf{D}_K$ be the mean value of $\mathbf{D}$ on the cell K, and we refer to Figure 3 for an illustration of the notation.

Assumption 2.6.

1. $\mathcal{D}$ is an admissible mesh.
2. For all boundary edge $b \in \mathcal{A}_{\text{ext}}$, denoting by $T(b)$ the control volume such that $b \in \mathcal{A}_{T(b)}$, we assume that the half-line starting at $X_{T(b)}$ and with direction $\mathbf{D}_{T(b)}\vec{n}_{T(b),b}$ intersects b at some point denoted by X_b.
3. For all $a \in \mathcal{A}_{\text{int}}$, we denote by $T_1(a)$ and $T_2(a)$ the control volumes on either side of a and we assume that, for $i = 1, 2$, the half-line starting at $X_{T_i(a)}$ and with

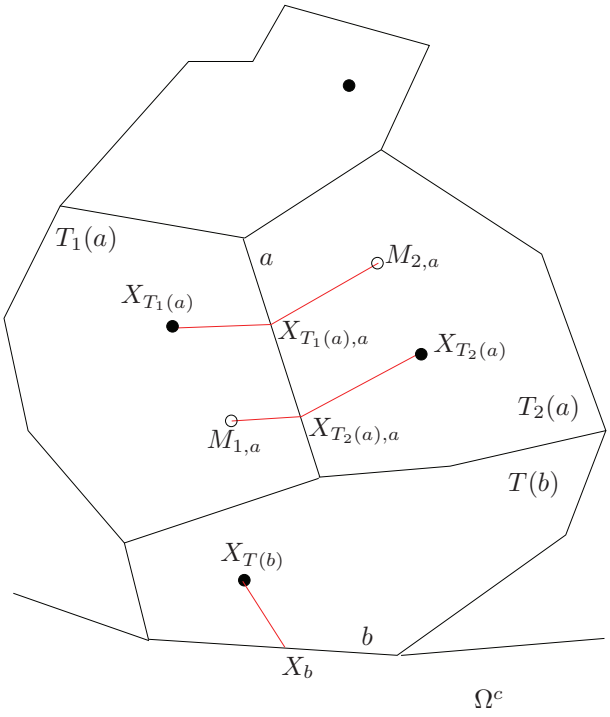

FIG. 3. *Illustration of Assumption* 2.6 *(choice of* $M_{i,a}$ *in the heterogeneous case).*

direction $\mathbf{D}_{T_i(a)}\vec{n}_{T_i(a),a}$ intersects a at some point denoted by $X_{T_i(a),a}$. We then take two points $M_{1,a}$ and $M_{2,a}$ inside the convex hull of the cell centers and boundary points $((X_K)_{K\in\mathcal{T}}, (X_b)_{b\in\mathcal{A}_{\rm ext}})$ such that

(i) $M_{2,a}$ belongs to the (open) half-line starting at $X_{T_1(a),a}$ and with direction $\mathbf{D}_{T_2(a)}\vec{n}_{T_1(a),a}$;

(ii) $M_{1,a}$ belongs to the (open) half-line starting at $X_{T_2(a),a}$ and with direction $\mathbf{D}_{T_1(a)}\vec{n}_{T_2(a),a}$.

We let $\mathcal{M}$ be the set of the chosen points $(M_{i,a})_{a\in\mathcal{A}_{\rm int},\, i=1,2}$.

Remark 2.7. The half-line starting at $X_{T_i(a)}$ and with direction $\mathbf{D}_{T_i(a)}\vec{n}_{T_i(a),a}$ does not really need to intersect a: if this is not the case, we define $X_{T_i(a),a}$ as the intersection of this half-line with the hyperplane containing a. The coercivity assumption on $\mathbf{D}$ ensures that $X_{T_i(a),a}$ stays within distance of order $\mathrm{diam}(T_i(a))$ of a, which is sufficient for the construction and study of the scheme.

The construction of the scheme follows the three steps presented in the isotropic homogeneous case.

Step 1 *(interpolation of additional approximate values of the solution).* This step is the same as in the isotropic homogeneous case. For each interior edge a and $i = 1, 2$, we write

$$M_{i,a} = \alpha_{i,a} X_{T_i(a)} + \sum_{j=1}^{J_{i,a}} \lambda_{i,a}(j) X_{i,a}(j) \tag{16}$$

(where $(X_{i,a}(j))_{j=1,\dots,J_{i,a}}$ are cell centers or boundary points and the coefficients $(\alpha_{i,a}, (\lambda_{i,a}(j))_{j=1,\dots,J_{i,a}})$ define a convex combination with $\alpha_{i,a} > 0$). In practice, we

fix $J_{i,a} = 2$ and we search the closest points $(X_{i,a}(j))_{j=1,2}$ such that $M_{i,a}$ lays inside the triangular cell $(X_{T_i(a)}, X_{i,a}(1), X_{i,a}(2))$ (this search has a negligible computational cost); note that there is some freedom in the choice of $M_{i,a}$, that can be moved if a first choice does not belong to the convex hull of all the cell centers and boundary points.

Then, we define

$$u_{M_{i,a}} = \alpha_{i,a} u_{T_i(a)} + \sum_{j=1}^{J_{i,a}} \lambda_{i,a}(j) u_{X_{i,a}(j)} \tag{17}$$

with $u_{X_{i,a}(j)} = u_K$ if $X_{i,a}(j) = X_K$ with $K \in \mathcal{T}$ and $u_{X_{i,a}(j)} = u_b$ if $X_{i,a}(j) = X_b$ with $b \in \mathcal{A}_{\text{ext}}$.

Remark 2.8. We do *not* yet define some approximate values of the solution at the points $X_{T_i(a),a}$: such values will be imposed by the conservativity of the fluxes.

Step 2 *(definition of the approximate value of the flux $\vec{q}\cdot\vec{n}_{K,a}$).* Let K be a cell and $a \in \mathcal{A}_K$. If a is a boundary edge and $\vec{n}_a$ is any unit normal to a, a consistent approximation of the flux of $\vec{q} = -\mathbf{D}\nabla u$ through a outside the control volume K is given by

$$F_{K,a}(u) = |\mathbf{D}_K \vec{n}_a| \, \frac{u_K - u_a}{d(X_a, X_K)}. \tag{18}$$

If a is an interior edge, denoting by $u_{T_i(a),a}$ an approximate value (not yet known) of the solution at $X_{T_i(a),a}$, we have four consistent ways, $F^1_{1,a}(u)$, $F^2_{1,a}(u)$, $F^1_{2,a}(u)$, and $F^2_{2,a}(u)$, to compute the flux of $\vec{q}$ going through a (the first two corresponding to fluxes outside $T_1(a)$, the last two to fluxes outside $T_2(a)$):

$$\begin{aligned}
F^1_{1,a}(u) &= \left|\mathbf{D}_{T_1(a)}\vec{n}_a\right| \frac{u_{T_1(a)} - u_{T_1(a),a}}{d(X_{T_1(a)}, X_{T_1(a),a})}, \quad F^2_{1,a}(u) = \left|\mathbf{D}_{T_2(a)}\vec{n}_a\right| \frac{u_{T_1(a),a} - u_{M_{2,a}}}{d(X_{T_1(a),a}, M_{2,a})}, \\
F^1_{2,a}(u) &= \left|\mathbf{D}_{T_2(a)}\vec{n}_a\right| \frac{u_{T_2(a)} - u_{T_2(a),a}}{d(X_{T_2(a)}, X_{T_2(a),a})}, \quad F^2_{2,a}(u) = \left|\mathbf{D}_{T_1(a)}\vec{n}_a\right| \frac{u_{T_2(a),a} - u_{M_{1,a}}}{d(X_{T_2(a),a}, M_{1,a})}.
\end{aligned} \tag{19}$$

Imposing the conservativity relations $F^1_{1,a}(u) = F^2_{1,a}(u)$ and $F^1_{2,a}(u) = F^2_{2,a}(u)$ allows us to compute the edge values as convex combinations of $(u_{T_i(a)})_{i=1,2}$ and $(u_{M_{i,a}})_{i=1,2}$:

$$\begin{aligned}
u_{T_1(a),a} &= \frac{\frac{\left|\mathbf{D}_{T_2(a)}\vec{n}_a\right|}{d(X_{T_1(a),a}, M_{2,a})} u_{M_{2,a}} + \frac{\left|\mathbf{D}_{T_1(a)}\vec{n}_a\right|}{d(X_{T_1(a)}, X_{T_1(a),a})} u_{T_1(a)}}{\frac{\left|\mathbf{D}_{T_2(a)}\vec{n}_a\right|}{d(X_{T_1(a),a}, M_{2,a})} + \frac{\left|\mathbf{D}_{T_1(a)}\vec{n}_a\right|}{d(X_{T_1(a)}, X_{T_1(a),a})}}, \\
u_{T_2(a),a} &= \frac{\frac{\left|\mathbf{D}_{T_1(a)}\vec{n}_a\right|}{d(X_{T_2(a),a}, M_{1,a})} u_{M_{1,a}} + \frac{\left|\mathbf{D}_{T_2(a)}\vec{n}_a\right|}{d(X_{T_2(a)}, X_{T_2(a),a})} u_{T_2(a)}}{\frac{\left|\mathbf{D}_{T_1(a)}\vec{n}_a\right|}{d(X_{T_2(a),a}, M_{1,a})} + \frac{\left|\mathbf{D}_{T_2(a)}\vec{n}_a\right|}{d(X_{T_2(a)}, X_{T_2(a),a})}}.
\end{aligned} \tag{20}$$

Defining

$$\begin{aligned}
\delta_{1,a} &= \frac{d(X_{T_2(a),a}, M_{1,a})}{\left|\mathbf{D}_{T_1(a)}\vec{n}_a\right|} + \frac{d(X_{T_2(a)}, X_{T_2(a),a})}{\left|\mathbf{D}_{T_2(a)}\vec{n}_a\right|}, \\
\delta_{2,a} &= \frac{d(X_{T_1(a),a}, M_{2,a})}{\left|\mathbf{D}_{T_2(a)}\vec{n}_a\right|} + \frac{d(X_{T_1(a)}, X_{T_1(a),a})}{\left|\mathbf{D}_{T_1(a)}\vec{n}_a\right|},
\end{aligned} \tag{21}$$

the fluxes outside $T_1(a)$ and outside $T_2(a)$ then have the following expressions:

$$F_{1,a}(u) = \frac{u_{T_1(a)} - u_{M_{2,a}}}{\delta_{2,a}} \quad \text{and} \quad F_{2,a}(u) = \frac{u_{T_2(a)} - u_{M_{1,a}}}{\delta_{1,a}}. \tag{22}$$

Using (17), this leads to

$$\begin{aligned} F_{1,a}(u) &= \frac{\alpha_{2,a}(u_{T_1(a)} - u_{T_2(a)}) + \sum_{j=1}^{J_{2,a}} \lambda_{2,a}(j)(u_{T_1(a)} - u_{X_{2,a}(j)})}{\delta_{2,a}}, \\ F_{2,a}(u) &= \frac{\alpha_{1,a}(u_{T_2(a)} - u_{T_1(a)}) + \sum_{j=1}^{J_{1,a}} \lambda_{1,a}(j)(u_{T_2(a)} - u_{X_{1,a}(j)})}{\delta_{1,a}}. \end{aligned} \tag{23}$$

The rest is identical to the isotropic homogeneous case (the anisotropy and heterogeneity have been taken into account in $\delta_{1,a}$ and $\delta_{2,a}$). We take

$$0 < \beta_{1,a} \le 2\frac{\alpha_{1,a}}{\delta_{1,a}} \quad \text{and} \quad 0 < \beta_{2,a} \le 2\frac{\alpha_{2,a}}{\delta_{2,a}}, \tag{24}$$

we define

$$G_{1,a}(u) = F_{1,a}(u) - \beta_{2,a}(u_{T_1(a)} - u_{T_2(a)}),\ G_{2,a}(u) = F_{2,a}(u) - \beta_{1,a}(u_{T_2(a)} - u_{T_1(a)}),$$

and we let

(25)

(i) if $|G_{1,a}(u)| + |G_{2,a}(u)| \neq 0$:

$$\gamma_{1,a}(u) = \frac{|G_{2,a}(u)|}{|G_{1,a}(u)| + |G_{2,a}(u)|} \quad \text{and} \quad \gamma_{2,a}(u) = \frac{|G_{1,a}(u)|}{|G_{1,a}(u)| + |G_{2,a}(u)|},$$

(ii) else $\gamma_{1,a}(u) = \gamma_{2,a}(u) = \frac{1}{2}$.

Letting $\epsilon_{i,K} = +1$ if $K = T_i(a)$ and $\epsilon_{i,K} = -1$ otherwise, the approximation of $\vec{q}\cdot\vec{n}_{K,a}$ is then

$$F_{K,a}(u) = \gamma_{1,a}(u)\epsilon_{1,K}F_{1,a}(u) + \gamma_{2,a}(u)\epsilon_{2,K}F_{2,a}(u). \tag{26}$$

Step 3 *(the scheme).* The scheme is the same as (14): find $u = (u_K)_{K\in\mathcal{T}}$ such that

$$\forall K \in \mathcal{T} : \sum_{a\in\mathcal{A}_K} |a| F_{K,a}(u) = \int_K f. \tag{27}$$

Remark 2.9 (case of triangular cells in dimension $d = 2$). We refer to [33] for a description of possible choices of the cell centers and the other parameters of the scheme in the case of a triangular mesh in dimension $d = 2$.

3. Study of the scheme. Assumption 2.6 being satisfied, we denote by $\mathcal{S}$ the scheme (27), where the interior fluxes are given by (26) and the boundary fluxes are given by (18), with the notation (16), (17), (21), (23), and (25) and the choice (24).

In this section, we provide some theoretical results on $\mathcal{S}$, starting by recalling the following proposition that has been proved during the construction of the scheme.

PROPOSITION 3.1. *The scheme $\mathcal{S}$ has the LMP structure.*

Remark 3.2. For most meshes, if $T_i(a)$ is an interior cell (i.e., such that no edge of $T_i(a)$ belongs to $\mathcal{A}_{\text{ext}}$), it is possible to choose the convex combination (16) such that all the $X_{i,a}(j)$ are cell centers and not boundary points; in this case, boundary values can be involved in the definition (26) of $F_{K,a}(u)$ only if K is a boundary cell or the neighbor of a boundary cell (i.e., K belongs to the "first and second layers" of cells from $\partial\Omega$). Proposition 1.4 then shows that, for a nonnegative right-hand side and homogeneous boundary conditions, the solution to the scheme attains its minimum in a cell within the first or second layers from $\partial\Omega$.

If one manages to write all the convex combinations (16) using only cell centers, then (26) involves boundary values only if K is a boundary cell (i.e., $\mathcal{A}_K \cap \mathcal{A}_{\text{ext}} \neq \emptyset$); in this case, for $f \geq 0$ and $u_\partial = 0$, the solution to the scheme cannot attain is minimum on an interior cell unless it is constant.

3.1. Convergence. To simplify the notation and reasoning, we assume in this section that the boundary conditions are homogeneous: $u_\partial = 0$.

Let $S_\mathcal{D} : \mathbb{R}^{\text{Card}(\mathcal{T})} \mapsto \mathbb{R}^{\text{Card}(\mathcal{T})}$ be defined by $S_\mathcal{D}(u)_K = -\sum_{a\in\mathcal{A}_K} |a| F_{K,a}(u)$. Finding a solution to the scheme $\mathcal{S}$ consists of solving in $u = (u_K)_{K\in\mathcal{T}}$ the nonlinear equation $S_\mathcal{D}(u) = (\int_K f)_{K\in\mathcal{T}}$; a usual requirement when solving such an equation, especially if one wants a numerically stable solution, is the continuity of $S_\mathcal{D}$. Under the general framework we consider above, this continuity is not obvious since $\gamma_{i,a}(u)$ given by (25) can be noncontinuous. One way to solve this issue is to impose that, if the denominator involved in the definition of $\gamma_{i,a}(u)$ tends to zero, then both fluxes $\epsilon_{1,K}F_{1,a}(u)$ and $\epsilon_{2,K}F_{2,a}(u)$ tend to the same value; this can be done by choosing the $\beta_{i,a}$ (depending only on the mesh) in (24) such that

$$\forall a \in \mathcal{A}_{\text{int}} \,:\, \beta_{1,a} = \beta_{2,a}. \tag{28}$$

Indeed, in this case, if

$$\begin{aligned} &|G_{1,a}(u)| + |G_{2,a}(u)| \\ &\quad = |F_{1,a}(u) - \beta_{2,a}(u_{T_1(a)} - u_{T_2(a)})| + |F_{2,a}(u) - \beta_{1,a}(u_{T_2(a)} - u_{T_1(a)})| \to 0 \end{aligned}$$

as u tends to some $\widetilde{u}$, then $\epsilon_{1,K}F_{1,a}(u)$ and $\epsilon_{2,K}F_{2,a}(u)$ both tend to the same quantity $\epsilon_{1,K}\beta_{2,a}(\widetilde{u}_{T_1(a)} - \widetilde{u}_{T_2(a)}) = \epsilon_{2,K}\beta_{1,a}(\widetilde{u}_{T_2(a)} - \widetilde{u}_{T_1(a)}) = \epsilon_{1,K}F_{1,a}(\widetilde{u}) = \epsilon_{2,K}F_{2,a}(\widetilde{u})$ (recall that $\epsilon_{1,K} = -\epsilon_{2,K}$). Hence, whatever the values then taken by $\gamma_{1,a}(u)$ and $\gamma_{2,a}(u)$ as $u \to \widetilde{u}$, the convex combination $\gamma_{1,a}(u)\epsilon_{1,K}F_{1,a}(u) + \gamma_{2,a}(u)\epsilon_{2,K}F_{2,a}(u)$ tends to $\epsilon_{1,K}F_{1,a}(\widetilde{u}) = \epsilon_{2,K}F_{2,a}(\widetilde{u}) = \gamma_{1,a}(\widetilde{u})\epsilon_{1,K}F_{1,a}(\widetilde{u}) + \gamma_{2,a}(\widetilde{u})\epsilon_{2,K}F_{2,a}(\widetilde{u})$.

The proof of convergence of finite volume schemes for elliptic equations requires some discrete H^1_0 estimates on the solution. These estimates then give compactness properties that ensure that the solution to the scheme converges as the mesh size tends to 0 to a function in H^1_0; the reasoning is then concluded by proving that this function is the weak solution to the PDE.

These discrete H^1_0 estimates come from some coercivity property of the scheme, similar to the classical coercivity property of the bilinear form used in the variational formulation of (1). If the definition of the scheme is based on a scalar product, such as in the HMM or DDFV methods (see [20, 18]), then the coercivity property is immediate; in general, however, the maximum principle is lost in such methods. For schemes that first define fluxes and then write their balance, such as [15, 3] or the schemes we consider here, the coercivity property is not automatic (although numerically often satisfied), and it must be assumed during the theoretical study.

In the scheme $\mathcal{S}$, the fluxes $F_{K,a}(u)$ are nonlinear convex combinations of *linear* fluxes $F_{i,a}(u)$ (the nonlinearity being solely concentrated in the functions $\gamma_{i,a}(u)$); we take advantage of this special form to give a proof of convergence of $\mathcal{S}$ inspired from the technique developed in a linear setting in [3] (the main adaptations are in the separation of the linear and nonlinear parts of the fluxes, as well as in the consideration that some of the fluxes can be nonconsistent—see Lemma 3.5 and section 3.1.3. The coercivity assumption we adopt is therefore close to the one selected in [3].

If $K \in \mathcal{T}$ and $a \in \mathcal{A}_K$, we denote by $d_{K,a}$ the orthogonal distance between X_K and the hyperplane containing a; we define $d_a = d_{T_1(a),a} + d_{T_2(a),a}$ if $a \in \mathcal{A}_{\rm int}$ and $d_a = d_{T(a),a}$ if $a \in \mathcal{A}_{\rm ext}$. The coercivity property we assume on the fluxes is

$$\begin{aligned} &\exists \zeta > 0 \text{ such that, } \forall v = (v_K)_{K\in\mathcal{T}} : \\ &\quad \sum_{a\in\mathcal{A}_{\rm int}} |a| \min\left[\epsilon_{1,K} F_{1,a}(v)(v_K - v_L); \epsilon_{2,K} F_{2,a}(v)(v_K - v_L)\right] \\ &\quad + \sum_{a\in\mathcal{A}_{\rm ext}} |a| F_{K,a}(v) v_K \geq \zeta \left(\sum_{a\in\mathcal{A}_{\rm int}} \frac{|a|}{d_a}(v_K - v_L)^2 + \sum_{a\in\mathcal{A}_{\rm ext}} \frac{|a|}{d_a} v_K^2 \right), \end{aligned} \tag{29}$$

where, if $a \in \mathcal{A}_{\rm int}$, K and L denote the cells on each side of a, and, if $a \in \mathcal{A}_{\rm ext}$, K is the cell such that $a \in \mathcal{A}_K$. This relation basically demands that the linear fluxes $F_{1,a}(v)$ and $F_{2,a}(v)$ have a "strong enough" coefficient along $v_{T_1(a)} - v_{T_2(a)}$ (this quantity being the one used, when the mesh is admissible, in the sense of [22], to construct the two-point finite volume flux across a): one can heuristically consider that (29) asks that the coefficients $\alpha_{i,a}$ in (16) are large enough with respect to the coefficients $\lambda_{i,j}$ (see (23)). However, as mentioned above and as seen in [15, 3], for the kind of scheme we consider there does not seem to exist easily verifiable mathematical assumptions on the mesh, and the data that ensure that such a coercivity property holds. We nevertheless give in section 4 some comments on the general validity of (29), based on numerical computations.

It will be convenient to identify $v = (v_K)_{K\in\mathcal{T}} \in \mathbb{R}^{{\rm Card}(\mathcal{T})}$ with the piecewise-constant function v on Ω that takes the value v_K inside the cell K; we let $\mathcal{H}(\mathcal{T})$ be the space of such piecewise-constant functions.

The right-hand side in (29) is the discrete H_0^1 norm of interest to us; it will play an important role in the study of the scheme; thus with the same conventions as above, we introduce the notation

$$\|v\|_{\mathcal{D}}^2 = \sum_{a\in\mathcal{A}_{\rm int}} \frac{|a|}{d_a}(v_K - v_L)^2 + \sum_{a\in\mathcal{A}_{\rm ext}} \frac{|a|}{d_a} v_K^2. \tag{30}$$

3.1.1. A priori estimates. We need to introduce two quantities: the first one,

$${\rm reg}(\mathcal{D}) = \max_{(K,L)\in\mathcal{T}^2 \,|\, \mathcal{A}_K\cap\mathcal{A}_L\neq\emptyset} \frac{{\rm diam}(K)}{{\rm diam}(L)} + \max_{K\in\mathcal{T},\, a\in\mathcal{A}_K} \frac{{\rm diam}(K)}{d_{K,a}} + \max_{K\in\mathcal{T}} {\rm Card}(\mathcal{A}_K),$$

is a measure of the pure geometrical regularity of the mesh, and the second one,

$${\rm reg}_{\mathbf{D}}(\mathcal{D},\mathcal{M}) = {\rm reg}(\mathcal{D}) + \max_{a\in\mathcal{A}_{\rm int}} \left(\frac{{\rm diam}(T_1(a))}{d(X_{T_2(a),a}, M_{1,a})} + \frac{{\rm diam}(T_2(a))}{d(X_{T_1(a),a}, M_{2,a})} \right),$$

evaluates, on top of the geometrical regularity, the degree of compatibility with respect to $\mathbf{D}$ of the mesh and $\mathcal{M} = (M_{i,a})_{a\in\mathcal{A}_{\rm int},\, i=1,2}$ (points of interpolation of the approximate solution, coming from Assumption 2.6).

PROPOSITION 3.3 (a priori estimates). *Under Assumption* 2.6*, if* θ *is a real number such that* $a\theta \geq \mathrm{reg}_{\mathbf{D}}(\mathcal{D}, \mathcal{M})$ *and* (29) *is satisfied, then there exists* $C_1 > 0$ *depending only on* θ *and* ζ *such that, for any solution* $u \in \mathcal{H}(\mathcal{T})$ *to the scheme* $\mathcal{S}$*, we have* $||u||_{\mathcal{D}} \leq C_1 ||f||_{L^2(\Omega)}$.

Proof. Multiply (27) by u_K, sum on $K \in \mathcal{T}$, and use the conservativity of the fluxes to gather the sum on the edges. This gives

$$\int_\Omega fu = \sum_{K\in\mathcal{T}} \sum_{a\in\mathcal{A}_K} |a| F_{K,a}(u) u_K = \sum_{a\in\mathcal{A}} |a| F_{K,a}(u)(u_K - u_L),$$

where, in the last sum, if $a \in \mathcal{A}_{\mathrm{int}}$, then K and L are the two cells on each side of a, and, if $a \in \mathcal{A}_{\mathrm{ext}}$, K is the cell such that $a \in \mathcal{A}_K$ and $u_L = 0$. If a is an interior edge, since $F_{K,a}(u)(u_K - u_L)$ is a convex combination of $\epsilon_{1,K} F_{1,a}(u)(u_K - u_L)$ and $\epsilon_{2,K} F_{2,a}(u)(u_K - u_L)$, we have

$$F_{K,a}(u)(u_K - u_L) \geq \min\left[\epsilon_{1,K} F_{1,a}(u)(u_K - u_L); \epsilon_{2,K} F_{2,a}(u)(u_K - u_L)\right],$$

and, by (26) and (29), we infer

$$\int_\Omega fu \geq \zeta ||u||_{\mathcal{D}}^2. \tag{31}$$

From [19, proof of Lemma 6.2] or [23, Lemma 5.2], one immediately sees that there exists C_2 depending only on θ such that, for all $u \in \mathcal{H}(\mathcal{T})$,

$$||u||_{L^2(\Omega)} \leq C_2 ||u||_{\mathcal{D}} \tag{32}$$

(this is a discrete Poincaré's inequality). The conclusion of the proposition thus follows by applying the Cauchy–Schwarz inequality to the left-hand side of (31). ∎

3.1.2. Existence of a solution.

PROPOSITION 3.4 (existence of a solution to the scheme). *Under Assumption* 2.6*, if* (28) *and* (29) *are satisfied, then there exists at least one solution to the scheme* $\mathcal{S}$.

Proof. The proof is a straightforward application of Brouwer's topological degree (see [16]). We let $f_{\mathcal{T}} = (\int_K f)_{K\in\mathcal{T}} \in \mathcal{H}(\mathcal{T})$, and we show that there exists $R > 0$ such that, for all $t \in [0,1]$, the $L^2(\Omega)$ norm of any solution u to

$$tu + (1-t) S_{\mathcal{D}}(u) = f_{\mathcal{T}} \tag{33}$$

is bounded by R; this ensures that the degree of the continuous (see (28)) function $S_{\mathcal{D}}$ on the ball of radius R at $f_{\mathcal{T}}$ is equal to the degree of the identity mapping; upon increasing R, we can assume that the ball of radius R contains $f_{\mathcal{T}}$, and thus that this degree is equal to 1, which implies the existence of a solution to $S_{\mathcal{D}}(u) = f_{\mathcal{T}}$.

The desired a priori estimate on the solutions to (33) can be obtained exactly in the same way as in the proof of Proposition 3.3. Multiplying the coordinate K of the system (33) by u_K and summing on K we get, thanks to (29),

$$\int_\Omega fu \geq t \sum_{K\in\mathcal{T}} u_K^2 + (1-t)\zeta ||u||_{\mathcal{D}}^2.$$

Then using (32) and defining $m_{\mathcal{T}} = \min_{K\in\mathcal{T}}(1/|K|)$, we find

$$\min(m_{\mathcal{T}}, \zeta C_2^{-2}) ||u||_{L^2(\Omega)} \leq (t m_{\mathcal{T}} + (1-t)\zeta C_2^{-2}) ||u||_{L^2(\Omega)} \leq ||f||_{L^2(\Omega)}.$$

The value $R = ||f||_{L^2(\Omega)} / \min(m_{\mathcal{T}}, \zeta C_2^{-2})$ is therefore acceptable. ∎

3.1.3. Convergence of the scheme. We first prove that, under a natural assumption on the choice of $M_{i,a}$ (see Remark 2.3) and for edges across which $\mathbf{D}$ is continuous, the linear components of the flux (26) are consistent.

In the following lemma and after it, for any $\varphi \in C_c^2(\Omega)$, we denote by $\varphi_{\mathcal{D}} \in \mathcal{H}(\mathcal{T})$ the function $\varphi_{\mathcal{D}} = (\varphi(X_K))_{K\in\mathcal{T}}$. We also define the norm

$$||\varphi||_{C_c^2(\Omega)} = \sup_{x\in\Omega} \left(|\varphi(x)| + |\nabla\varphi(x)| + |D^2\varphi(x)|\right).$$

LEMMA 3.5 (consistency of the flux). *Let Assumption* 2.6 *hold, and let θ be a real number such that $\theta \geq \mathrm{reg}_{\mathbf{D}}(\mathcal{D},\mathcal{M})$. We also assume that*

$$\begin{aligned} &\exists \rho > 0 \text{ such that } \forall a \in \mathcal{A}_{\mathrm{int}}\,,\ \forall i = 1,2\,,\\ &\forall X \in \{X_{T_1(a)}, X_{T_2(a)}, X_{i,a}(j)\,;\ j = 1,\dots,J_{i,a}\}\ :\ d(M_{i,a},X) \leq \rho \mathrm{diam}(T_i(a)) \end{aligned} \tag{34}$$

(the $X_{i,a}(j)$ are the cell centers and boundary points chosen in (16)*). Then there exists C_3 depending only on θ, ρ, and $\mathbf{D}$ such that, for all $\varphi \in C_c^2(\Omega)$, all $K \in \mathcal{T}$, all $a \in \mathcal{A}_K$, and all $i = 1,2$, if $\mathbf{D}$ is Lipschitz-continuous on $T_1(a) \cup a \cup T_2(a)$,*

$$\left|\epsilon_{i,K} F_{i,a}(\varphi_{\mathcal{D}}) - \frac{1}{|K|}\int_K (-\mathbf{D}\nabla\varphi)\cdot\vec{n}_{K,a}\right| \leq C_3 ||\varphi||_{C_c^2(\Omega)} \mathrm{diam}(K).$$

Proof. To simplify the proof we let $i = 1$; the various $\mathcal{O}$ we use here depend only on θ, ρ, and $\mathbf{D}$ (i.e., $\mathcal{O}(z)$ is a quantity bounded, in absolute value, by $C_4|z|$ with C_4 depending only on θ, ρ, and $\mathbf{D}$).

Let us first check that the formula (20) for $\varphi_{T_1(a),a}$ is an order 2 approximation of $\varphi(X_{T_1(a),a})$ (with $\varphi_{M_{2,a}}$ computed by convex combination of $(\varphi(X_K))_{K\in\mathcal{T}}$ through (17)). Denoting by $\mathbf{D}_a$ the mean value of $\mathbf{D}$ on a ($\mathbf{D}$ is continuous across a), we have, by regularity of φ and Assumption (2.6),

$$\begin{aligned} \mathbf{D}_a\nabla\varphi(X_{T_1(a),a})\cdot\vec{n}_{T_1(a),a} = |\mathbf{D}_a\vec{n}_a|\,\frac{\varphi(X_{T_1(a),a}) - \varphi_{T_1(a)}}{d(X_{T_1(a),a}, X_{T_1(a)})}\\ + \mathcal{O}\left(||\varphi||_{C_c^2(\Omega)}\mathrm{diam}(T_1(a))\right) \end{aligned}$$

and

$$\begin{aligned} \mathbf{D}_a\nabla\varphi(X_{T_1(a),a})\cdot\vec{n}_{T_1(a),a} = |\mathbf{D}_a\vec{n}_a|\,\frac{\varphi(M_{2,a}) - \varphi(X_{T_1(a),a})}{d(M_{2,a}, X_{T_1(a),a})}\\ + \mathcal{O}\left(||\varphi||_{C_c^2(\Omega)}\mathrm{diam}(T_1(a))\right). \end{aligned}$$

The equality of the two right-hand sides and the Lipschitz-continuity of $\mathbf{D}$ on $T_1(a)\cup a\cup T_2(a)$ give

$$\begin{aligned} \left|\mathbf{D}_{T_1(a)}\vec{n}_a\right|\frac{\varphi(X_{T_1(a),a}) - \varphi_{T_1(a)}}{d(X_{T_1(a),a}, X_{T_1(a)})} = \left|\mathbf{D}_{T_2(a)}\vec{n}_a\right|\frac{\varphi(M_{2,a}) - \varphi(X_{T_1(a),a})}{d(M_{2,a}, X_{T_1(a),a})}\\ + \mathcal{O}\left(||\varphi||_{C_c^2(\Omega)}\mathrm{diam}(T_1(a))\right). \end{aligned}$$

But, by (16), (17), (34), and regularity of φ, we have

$$\varphi(M_{2,a}) = \varphi_{M_{2,a}} + \mathcal{O}\left(||\varphi||_{C_c^2(\Omega)}\mathrm{diam}(T_1(a))^2\right), \tag{35}$$

and thus, using $\theta \geq \mathrm{reg}_{\mathbf{D}}(\mathcal{D}, \mathcal{M})$,

$$\begin{aligned}\left|\mathbf{D}_{T_1(a)}\vec{n}_a\right| \frac{\varphi(X_{T_1(a),a}) - \varphi_{T_1(a)}}{d(X_{T_1(a),a}, X_{T_1(a)})} = \left|\mathbf{D}_{T_2(a)}\vec{n}_a\right| \frac{\varphi_{M_{2,a}} - \varphi(X_{T_1(a),a})}{d(M_{2,a}, X_{T_1(a),a})} \\ + \mathcal{O}\left(||\varphi||_{C_c^2(\Omega)}\mathrm{diam}(T_1(a))\right). \end{aligned} \tag{36}$$

Since

$$\frac{\left|\mathbf{D}_{T_2(a)}\vec{n}_a\right|}{d(M_{2,a}, X_{T_1(a),a})} + \frac{\left|\mathbf{D}_{T_1(a)}\vec{n}_a\right|}{d(X_{T_1(a),a}, X_{T_1(a)})} \geq \frac{C_5}{\mathrm{diam}(T_1(a))}$$

with C_5 depending only on $\mathbf{D}$ and θ, (36) and the definition (20) of $\varphi_{T_1(a),a}$ give

$$\varphi(X_{T_1(a),a}) = \varphi_{T_1(a),a} + \mathcal{O}\left(||\varphi||_{C_c^2(\Omega)}\mathrm{diam}(T_1(a))^2\right). \tag{37}$$

From the second order approximations (35) and (37) and formula (19), we clearly deduce

$$F^1_{1,a}(\varphi_{\mathcal{D}}) = (-\mathbf{D}_{T_1(a)}\nabla\varphi(X_{T_1(a),a}))\cdot\vec{n}_{T_1(a),a} + \mathcal{O}\left(||\varphi||_{C_c^2(\Omega)}\mathrm{diam}(T_1(a))\right)$$

and

$$F^2_{1,a}(\varphi_{\mathcal{D}}) = (-\mathbf{D}_{T_2(a)}\nabla\varphi(X_{T_1(a),a}))\cdot\vec{n}_{T_1(a),a} + \mathcal{O}\left(||\varphi||_{C_c^2(\Omega)}\mathrm{diam}(T_1(a))\right).$$

Recalling that $F_{1,a}(\varphi_{\mathcal{D}}) = F^1_{1,a}(\varphi_{\mathcal{D}}) = F^2_{1,a}(\varphi_{\mathcal{D}})$, the regularity of φ and the definition of $\mathrm{reg}_{\mathbf{D}}(\mathcal{D}, \mathcal{M})$ conclude the proof. ∎

As mentioned above, the proof of convergence relies on a compactness result (Lemma 3.7) that ensures that, up to a subsequence, the solution to the scheme converges to a function in H^1_0. The statement of this lemma requires the introduction of a discrete gradient. If $v \in \mathcal{H}(\mathcal{T})$, then we define

$$\begin{aligned}&\forall a \in \mathcal{A}_{\mathrm{int}} \,:\, v_a = \frac{v_{T_1(a)} + v_{T_2(a)}}{2}, \\ &\forall a \in \mathcal{A}_{\mathrm{ext}} \,:\, v_a = 0,\end{aligned} \tag{38}$$

and we define the discrete gradient of v as the piecewise-constant function $\nabla_{\mathcal{D}} v : \Omega \mapsto \mathbb{R}^N$ such that

$$\forall K \in \mathcal{T}\,,\ \forall x \in K \,:\, \nabla_{\mathcal{D}} v(x) = \frac{1}{|K|}\sum_{a\in\mathcal{A}_K} |a|(v_a - v_K)\vec{n}_{K,a}.$$

The Cauchy–Schwarz inequality and the fact that $\sum_{a\in\mathcal{A}_K} |a|d_{K,a} = N|K|$ (because $\frac{|a|d_{K,a}}{N}$ is the measure of the convex hull of $\{X_K\}\cup a$) show that

$$||\nabla_{\mathcal{D}} v||^2_{L^2(\Omega)^N} \leq N \sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K} \frac{|a|}{d_{K,a}}(v_a - v_K)^2. \tag{39}$$

Moreover, it is easy to see that, if $\theta \geq \mathrm{reg}(\mathcal{D})$, there exists C_6 depending only on θ such that

$$\sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K} \frac{|a|}{d_{K,a}}(v_K - v_a)^2 \leq C_6||v||^2_{\mathcal{D}}. \tag{40}$$

Remark 3.6. There is in fact a vast number of valid choices for the edge values v_a; any value that ensures that the L^2 norm of $\nabla_{\mathcal{D}} v$ is bounded from above by $||v||_{\mathcal{D}}$ can be used. This is, for example, the case, under the regularity assumptions on the mesh, for $v_a = v_{T_1(a),a}$, $v_a = v_{T_2(a),a}$ or any combination of these two values.

The following lemma is an immediate consequence of the definition of $\mathrm{reg}(\mathcal{D})$, (39), (40) and of the technique used in [23, Lemma 4.2] (with the help of the discrete Sobolev embeddings of [23, Lemma 5.2]; see also the proofs of [19, Lemmas 6.2 and 6.5] and [14, Lemma 4.4], for example).

LEMMA 3.7 (discrete compactness property). *Let $(\mathcal{D}^n)_{n\geq 1}$ be a family of admissible meshes such that $(\mathrm{reg}(\mathcal{D}^n))_{n\geq 1}$ is bounded and $\mathrm{size}(\mathcal{D}^n) \to 0$ as $n \to \infty$. If $v^n \in \mathcal{H}(\mathcal{T}^n)$ is such that $(||v^n||_{\mathcal{D}^n})_{n\geq 1}$ is bounded, then there exists $\bar{v} \in H_0^1(\Omega)$ such that, up to a subsequence as $n \to \infty$,*

(i) *$v^n \to \bar{v}$ strongly in $L^q(\Omega)$ for all $q < \frac{2N}{N-2}$;*

(ii) *$\nabla_{\mathcal{D}} v^n \to \nabla \bar{v}$ weakly in $L^2(\Omega)^N$.*

We can now state and prove the convergence theorem for $\mathcal{S}$.

THEOREM 3.8 (convergence of the scheme). *Let $(\mathcal{D}^n)_{n\geq 1}$ be a family of admissible meshes satisfying Assumption* 2.6 *with sets of points $(\mathcal{M}^n)_{n\geq 1}$ such that $(\mathrm{reg}_{\mathbf{D}}(\mathcal{D}^n, \mathcal{M}^n))_{n\geq 1}$ is bounded. We also assume that $\mathrm{size}(\mathcal{D}^n) \to 0$ as $n \to \infty$ and that, for each mesh in the family,* (34) *and* (29) *hold with ρ and ζ not depending on n. We let $u^n \in \mathcal{H}(\mathcal{T}^n)$ be a solution to the scheme $\mathcal{S}$ with $\mathcal{D} = \mathcal{D}^n$, and let $\bar{u} \in H_0^1(\Omega)$ be the weak solution to* (1).

Then, as $n \to \infty$, u^n converges to $\bar{u}$ in $L^q(\Omega)$ for all $q < \frac{2N}{N-2}$.

Proof. From Proposition 3.3 and Lemma 3.7, there exists $\bar{u} \in H_0^1(\Omega)$ such that, up to a subsequence, $u^n \to \bar{u}$ strongly in $L^q(\Omega)$ for all $q < \frac{2N}{N-2}$ and $\nabla_{\mathcal{D}} u^n \to \nabla \bar{u}$ weakly in $L^2(\Omega)^N$. If we prove that $\bar{u}$ is the weak solution to (1) then, this solution being unique, reasoning on all the subsequences of $(u^n)_{n\geq 1}$ gives the convergence of the whole sequence and concludes the proof.

To simplify the notation, we drop the index n and assume that $u = u^n$ converges to $\bar{u}$ as $\mathrm{size}(\mathcal{D}) \to 0$, and we prove that $\bar{u}$ is the weak solution to (1).

We denote, for $(v, w) \in \mathcal{H}(\mathcal{T})$,

$$\begin{array}{ll} \widetilde{F}_{K,a}(v,w) = \gamma_{1,a}(v)\epsilon_{1,K}F_{1,a}(w) + \gamma_{2,a}(v)\epsilon_{2,K}F_{2,a}(w) & \text{if } a \in \mathcal{A}_K \cap \mathcal{A}_{\mathrm{int}}, \\ \widetilde{F}_{K,a}(v,w) = F_{K,a}(w) & \text{if } a \in \mathcal{A}_K \cap \mathcal{A}_{\mathrm{ext}}. \end{array} \tag{41}$$

We have $F_{K,a}(v) = \widetilde{F}_{K,a}(v, v)$ (see (26)), $\widetilde{F}_{K,a}(v, w)$ is conservative (i.e., $\widetilde{F}_{K,a}(v, w) + \widetilde{F}_{L,a}(v, w) = 0$ if a is an edge between K and L), and, for all v, $\widetilde{F}_{K,a}(v, \cdot)$ is linear.

Let $\varphi \in C_c^\infty(\Omega)$, $\varphi_{\mathcal{D}} = (\varphi_K)_{K\in\mathcal{T}}$ with $\varphi_K = \varphi(X_K)$ and $v = u - \varphi_{\mathcal{D}}$. The definition (41) shows that

$$\widetilde{F}_{K,a}(u, v)(v_K - v_L) \geq \min\left[\epsilon_{1,K}F_{1,a}(v)(v_K - v_L); \epsilon_{2,K}F_{2,a}(v)(v_K - v_L)\right].$$

The coercivity assumption (29) thus gives

$$\begin{aligned} \zeta||u - \varphi_{\mathcal{D}}||_{\mathcal{D}}^2 &\leq \sum_{a\in\mathcal{A}} |a|\widetilde{F}_{K,a}(u, u - \varphi_{\mathcal{D}})(v_K - v_L) \\ &\leq \sum_{a\in\mathcal{A}} |a|\widetilde{F}_{K,a}(u, u)(v_K - v_L) - \sum_{a\in\mathcal{A}} |a|\widetilde{F}_{K,a}(u, \varphi_{\mathcal{D}})(v_K - v_L). \end{aligned}$$

Defining $(u_a)_{a\in\mathcal{A}}$ and $(\varphi_a)_{a\in\mathcal{A}}$ (and thus $v_a = u_a - \varphi_a$) from u and $\varphi_{\mathcal{D}}$ by (38) and

using the conservativity of the fluxes, we obtain, by the balance equation (27),

$$\begin{aligned}\zeta||u-\varphi_{\mathcal{D}}||_{\mathcal{D}}^2 &\le \sum_{a\in\mathcal{A}}|a|F_{K,a}(u)(v_K-v_L)-\sum_{a\in\mathcal{A}}|a|\widetilde{F}_{K,a}(u,\varphi_{\mathcal{D}})(v_K-v_a-(v_L-v_a))\\ &\le \sum_{K\in\mathcal{T}}\left(\sum_{a\in\mathcal{A}_K}|a|F_{K,a}(u)\right)v_K-\sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K}|a|\widetilde{F}_{K,a}(u,\varphi_{\mathcal{D}})(v_K-v_a)\\ &\le \int_\Omega f(u-\varphi_{\mathcal{D}})-Q_1^{\mathcal{D}}. \qquad (42)\end{aligned}$$

Let us define $R_{K,a}=\widetilde{F}_{K,a}(u,\varphi_{\mathcal{D}})-\frac{1}{|K|}\int_K(-\mathbf{D}\nabla\varphi)\cdot\vec{n}_{K,a}$. We can write

$$\begin{aligned}Q_1^{\mathcal{D}} &= \sum_{K\in\mathcal{T}}\int_K(-\mathbf{D}\nabla\varphi)\cdot\frac{1}{|K|}\sum_{a\in\mathcal{A}_K}|a|(v_K-v_a)\vec{n}_{K,a}+\sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K}|a|R_{K,a}(v_K-v_a)\\ &= \int_\Omega\mathbf{D}\nabla\varphi\cdot\nabla_{\mathcal{D}}(u-\varphi_{\mathcal{D}})+\sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K}|a|R_{K,a}(v_K-v_a).\end{aligned}$$

Therefore, thanks to the Cauchy–Schwarz inequality and (40),

$$\left|Q_1^{\mathcal{D}}-\int_\Omega\mathbf{D}\nabla\varphi\cdot\nabla_{\mathcal{D}}(u-\varphi_{\mathcal{D}})\right|\le\left(\sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K}|a|d_{K,a}R_{K,a}^2\right)^{1/2}C_6^{1/2}||u-\varphi_{\mathcal{D}}||_{\mathcal{D}}.$$

The regularity of φ (as well as the boundedness of $\mathrm{reg}_{\mathbf{D}}(\mathcal{D},\mathcal{M})$) clearly shows that $||\varphi_{\mathcal{D}}||_{\mathcal{D}}$ is bounded as $\mathrm{size}(\mathcal{D})\to 0$; applying Lemma 3.7, and since $\varphi_{\mathcal{D}}\to\varphi$, we see that $\nabla_{\mathcal{D}}\varphi_{\mathcal{D}}\to\nabla\varphi$ weakly in $L^2(\Omega)^N$; hence, $||u||_{\mathcal{D}}$ remaining bounded and $\nabla_{\mathcal{D}}u$ weakly converging to $\nabla\bar{u}$ in $L^2(\Omega)^N$, we obtain C_7 such that

(43)

$$\limsup_{\mathrm{size}(\mathcal{D})\to 0}\left|Q_1^{\mathcal{D}}-\int_\Omega\mathbf{D}\nabla\varphi\cdot\nabla(\bar{u}-\varphi)\right|\le C_7\limsup_{\mathrm{size}(\mathcal{D})\to 0}\left(\sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K}|a|d_{K,a}R_{K,a}^2\right)^{1/2}.$$

By Lemma 3.5 and the definition of $R_{K,a}$, there exists C_8 not depending on the mesh such that, if $T_1(a)\cup a\cup T_2(a)$ does not intersect a discontinuity of $\mathbf{D}$, then $|R_{K,a}|\le C_8\mathrm{size}(\mathcal{D})$. If $\mathbf{D}$ is not continuous on $T_1(a)\cup a\cup T_2(a)$, then by regularity of φ we just have $|R_{K,a}|\le C_9$ for some C_9 not depending on the mesh. Let $\mathcal{A}_{\mathbf{D}}=\{a\in\mathcal{A}\mid\mathbf{D}\text{ is not continuous on }T_1(a)\cup a\cup T_2(a)\}$; we have

$$\begin{aligned}\sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K}|a|d_{K,a}R_{K,a}^2 &= \sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K\setminus\mathcal{A}_{\mathbf{D}}}|a|d_{K,a}R_{K,a}^2+\sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K\cap\mathcal{A}_{\mathbf{D}}}|a|d_{K,a}R_{K,a}^2\\ &\le C_8^2\mathrm{size}(\mathcal{D})^2N|\Omega|+C_9^2\sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K\cap\mathcal{A}_{\mathbf{D}}}|a|d_{K,a}. \qquad (44)\end{aligned}$$

Let Λ be the set of discontinuities of $\mathbf{D}$ (Λ does not depend on the mesh). The quantity

$$\frac{1}{N}\sum_{K\in\mathcal{T}}\sum_{a\in\mathcal{A}_K\cap\mathcal{A}_{\mathbf{D}}}|a|d_{K,a}$$

is the sum of the measures of the disjoint convex hulls of $\{X_K\} \cup a$ such that $a \in \mathcal{A}_{\mathbf{D}}$. But any edge a belonging to $\mathcal{A}_{\mathbf{D}}$ lays within distance size($\mathcal{D}$) of Λ, and the convex hull of $\{X_K\} \cup a$ is thus contained in $\Lambda_{2\text{size}(\mathcal{D})} = \{x \in \Omega \mid d(x, \Lambda) \leq 2\text{size}(\mathcal{D})\}$. Hence,

$$\sum_{K \in \mathcal{T}} \sum_{a \in \mathcal{A}_K \cap \mathcal{A}_{\mathbf{D}}} |a| d_{K,a} \leq N |\Lambda_{2\text{size}(\mathcal{D})}|$$

and, coming back to (44), we obtain

$$\sum_{K \in \mathcal{T}} \sum_{a \in \mathcal{A}_K} |a| d_{K,a} R_{K,a}^2 \leq C_8^2 \text{size}(\mathcal{D})^2 N |\Omega| + C_9^2 N |\Lambda_{2\text{size}(\mathcal{D})}|.$$

Since $\mathbf{D}$ is piecewise Lipschitz-continuous, Λ is a finite union of $(N-1)$-dimensional manifolds, and thus $|\Lambda_{2\text{size}(\mathcal{D})}|$ tends to 0 as size($\mathcal{D}$) $\to 0$; we deduce that

$$\sum_{K \in \mathcal{T}} \sum_{a \in \mathcal{A}_K} |a| d_{K,a} R_{K,a}^2 \to 0 \quad \text{as size}(\mathcal{D}) \to 0. \tag{45}$$

Using this in (43), we obtain

$$Q_1^{\mathcal{D}} \to \int_\Omega \mathbf{D} \nabla \varphi \cdot \nabla (\bar{u} - \varphi).$$

Hence, we can pass to the limit in (42) to deduce

$$\limsup_{\text{size}(\mathcal{D}) \to 0} ||u - \varphi_{\mathcal{D}}||_{\mathcal{D}}^2 \leq \frac{1}{\zeta} \left(\int_\Omega f(\bar{u} - \varphi) - \int_\Omega \mathbf{D} \nabla \varphi \cdot \nabla (\bar{u} - \varphi) \right). \tag{46}$$

Inequalities (39) and (40) show that $||\nabla_{\mathcal{D}} u - \nabla_{\mathcal{D}} \varphi_{\mathcal{D}}||_{L^2(\Omega)^N}^2 \leq N C_6 ||u - \varphi_{\mathcal{D}}||_{\mathcal{D}}^2$, and the weak convergence in $L^2(\Omega)^N$ of $\nabla_{\mathcal{D}} u - \nabla_{\mathcal{D}} \varphi_{\mathcal{D}}$ toward $\nabla \bar{u} - \nabla \varphi$ therefore leads to

$$||\nabla \bar{u} - \nabla \varphi||_{L^2(\Omega)^N}^2 \leq \frac{N C_6}{\zeta} \left(\int_\Omega f(\bar{u} - \varphi) - \int_\Omega \mathbf{D} \nabla \varphi \cdot \nabla (\bar{u} - \varphi) \right).$$

Letting $\varphi \in C_c^\infty(\Omega)$ tend strongly in $H_0^1(\Omega)$ to the weak solution $\widetilde{u}$ of (1) shows that this inequality also holds with $\varphi = \widetilde{u}$, in which case the right-hand side is equal to zero (precisely because $\widetilde{u}$ is the solution to (1)). We conclude that $||\nabla \bar{u} - \nabla \widetilde{u}||_{L^2(\Omega)^N} = 0$, that is, $\bar{u} = \widetilde{u}$, and the proof is complete. □

Following [3], it is also possible to reconstruct from the fluxes a discrete gradient that strongly converges toward the gradient of the continuous solution. For any $a \in \mathcal{A}$, we let $\bar{x}_a$ be the center of gravity of a. If $v \in \mathcal{H}(\mathcal{T})$, we define the discrete gradient $\overline{\nabla}_{\mathcal{D}}(v) : \Omega \to \mathbb{R}^N$ by

$$\forall K \in \mathcal{T}\,,\ \forall x \in K\ :\ \overline{\nabla}_{\mathcal{D}}(v)(x) = \frac{1}{|K|} \mathbf{D}_K^{-1} \sum_{a \in \mathcal{A}_K} |a| F_{K,a}(v)(X_K - \bar{x}_a),$$

where $F_{K,a}(v)$ is given by (26). Note that this discrete gradient $\overline{\nabla}_{\mathcal{D}}$ is *not* a linear operator; however, it has clearly identified linear and nonlinear parts: we will use this in the proof of its strong convergence.

This proof requires a slightly more restrictive assumption than (34) on the choice of the points used in the convex combinations (16):

$$\begin{aligned} &\exists \xi > 0\,,\ \exists E \in \mathbb{N} \text{ s.t. } \forall a \in \mathcal{A}_{\text{int}}\,,\ \forall i = 1, 2\,,\ \forall j = 1, \ldots, J_{i,a}\,, \text{ there exist} \\ &\text{continuous paths between } X_{i,a}(j) \text{ and } X_{T_1(a)},\ X_{T_2(a)} \text{ that cross} \\ &\text{at most } E \text{ edges, all having an } (N-1)\text{-dimensional measure in } [\tfrac{1}{\xi}|a|, \xi|a|]. \end{aligned} \tag{47}$$

This assumption is in fact only formally stronger than (34): in practical situations, the points involved in (16) are chosen within at most one or two cells of $T_1(a)$ and $T_2(a)$, and both (34) and (47) are satisfied.

THEOREM 3.9 (strong convergence of the gradient). *Under the assumptions of Theorem* 3.8*, if all the meshes satisfy* (47) *for some* ξ *and* E *not depending on* n*, then* $\overline{\nabla}_{\mathcal{D}^n}(u^n) \to \nabla\bar{u}$ *strongly in* $L^2(\Omega)^N$ *as* $n \to \infty$.

Proof. We drop the index n, and for $(v,w) \in \mathcal{H}(\mathcal{T})$ we define $\overline{\nabla}_{\mathcal{D},v} w : \Omega \to \mathbb{R}^N$ by

$$\forall K \in \mathcal{T},\ \forall x \in K\ :\ \overline{\nabla}_{\mathcal{D},v} w(x) = \frac{1}{|K|}\mathbf{D}_K^{-1} \sum_{a \in \mathcal{A}_K} |a| \widetilde{F}_{K,a}(v,w)(X_K - \bar{x}_a), \tag{48}$$

where $\widetilde{F}_{K,a}$ is given by (41). $\overline{\nabla}_{\mathcal{D},v}$ is a linear operator and $\overline{\nabla}_{\mathcal{D},v}(v) = \overline{\nabla}_{\mathcal{D}}(v)$. By (47), a given cell center or boundary point can appear in (16) only if it is within at most E cells of a; since $\mathrm{reg}_{\mathbf{D}}(\mathcal{D},\mathcal{M})$ stays bounded, this shows that each cell center or boundary point appears in the convex combinations (16) for at most C_{10} edges a, with C_{10} not depending on the mesh. Hence, from the expressions (22) of $F_{i,a}(w)$, the definitions (17) of $w_{M_{i,a}}$, and (47), we obtain C_{11} not depending on the mesh or (v,w) such that

$$||\overline{\nabla}_{\mathcal{D},v} w||_{L^2(\Omega)^N} \le C_{11}||w||_{\mathcal{D}}. \tag{49}$$

Let $\varepsilon > 0$, and fix $\varphi \in C_c^\infty(\Omega)$ within distance ε of $\bar{u}$ in $H_0^1(\Omega)$. We have

$$\begin{aligned} ||\overline{\nabla}_{\mathcal{D}}(u) - \nabla\bar{u}||_{L^2(\Omega)^N} &\le ||\overline{\nabla}_{\mathcal{D},u} u - \overline{\nabla}_{\mathcal{D},u}\varphi_{\mathcal{D}}||_{L^2(\Omega)^N} + ||\overline{\nabla}_{\mathcal{D},u}\varphi_{\mathcal{D}} - \nabla\varphi||_{L^2(\Omega)^N} + \varepsilon \\ &\le C_{11}||u - \varphi_{\mathcal{D}}||_{\mathcal{D}} + ||\overline{\nabla}_{\mathcal{D},u}\varphi_{\mathcal{D}} - \nabla\varphi||_{L^2(\Omega)^N} + \varepsilon. \end{aligned}$$

Using (46) and the fact that $||\varphi - \bar{u}||_{H_0^1(\Omega)} \le \varepsilon$, we can write, if size$(\mathcal{D})$ is small enough,

$$||\overline{\nabla}_{\mathcal{D}}(u) - \nabla\bar{u}||_{L^2(\Omega)^N} \le 2\varepsilon + ||\overline{\nabla}_{\mathcal{D},u}\varphi_{\mathcal{D}} - \nabla\varphi||_{L^2(\Omega)^N}.$$

It remains to prove that $\overline{\nabla}_{\mathcal{D},u}$ is strongly consistent, i.e., that $||\overline{\nabla}_{\mathcal{D},u}\varphi_{\mathcal{D}} - \nabla\varphi||_{L^2(\Omega)^N}$ tends to 0; this is a consequence of Lemma 3.5 and of the use of $R_{K,a}$ from the proof of Theorem 3.8.

Let us define $S_{K,a} = \widetilde{F}_{K,a}(u,\varphi_{\mathcal{D}}) - (-\mathbf{D}_K(\nabla\varphi)_K)\cdot\vec{n}_{K,a}$, where we let $(\nabla\varphi)_K = \frac{1}{|K|}\int_K \nabla\varphi$; we have, for all $x \in K$ and invoking [19, Lemma 6.1],

$$\begin{aligned} \overline{\nabla}_{\mathcal{D},u}\varphi_{\mathcal{D}}(x) &= \frac{1}{|K|}\sum_{a\in\mathcal{A}_K} |a|(\nabla\varphi)_K\cdot\vec{n}_{K,a}(\bar{x}_a - X_K) \\ &\quad + \frac{1}{|K|}\mathbf{D}_K^{-1}\sum_{a\in\mathcal{A}_K}|a|S_{K,a}(X_K - \bar{x}_a) \\ &= (\nabla\varphi)_K + \frac{1}{|K|}\mathbf{D}_K^{-1}\sum_{a\in\mathcal{A}_K}|a|S_{K,a}(X_K - \bar{x}_a). \end{aligned}$$

Hence, from the boundedness of $\mathrm{reg}_{\mathbf{D}}(\mathcal{D},\mathcal{M})$ we obtain C_{12} not depending on the

mesh such that

$$||\overline{\nabla}_{\mathcal{D},u}\varphi_{\mathcal{D}} - \nabla\varphi||^2_{L^2(\Omega)^N} \leq C_{12} \sum_{K\in\mathcal{T}} \frac{1}{|K|} \left(\sum_{a\in\mathcal{A}_K} |a|\, |S_{K,a}| d_{K,a} \right)^2$$
$$\leq C_{12} \sum_{K\in\mathcal{T}} \frac{1}{|K|} \left(\sum_{a\in\mathcal{A}_K} |a| d_{K,a} \right) \left(\sum_{a\in\mathcal{A}_K} |a| d_{K,a} S^2_{K,a} \right)$$
$$\leq C_{12} N \sum_{K\in\mathcal{T}} \sum_{a\in\mathcal{A}_K} |a| d_{K,a} S^2_{K,a}.$$

But, defining $R_{K,a}$ as in the proof of Theorem 3.8, the regularity of φ ensures that $|S_{K,a}| \leq |R_{K,a}| + C_{13}\text{size}(\mathcal{D})$ with C_{13} depending only on $\mathbf{D}$ and φ. Thus,

$$||\overline{\nabla}_{\mathcal{D},u}\varphi_{\mathcal{D}} - \nabla\varphi||^2_{L^2(\Omega)^N} \leq 2C_{12}N \sum_{K\in\mathcal{T}} \sum_{a\in\mathcal{A}_K} |a| d_{K,a} R^2_{K,a} + 2C_{12}N^2|\Omega| C^2_{13}\text{size}(\mathcal{D})^2,$$

and (45) concludes the proof. ∎

Remark 3.10. Using the same technique, it is very easy to prove the strong convergence of other kinds of discrete gradients; the important tools are an estimate of the kind (49) (the only reason why we introduced hypothesis (47)) and a consistency property. For example, the strong convergence also holds for the *linear* discrete gradients $\overline{\nabla}_{\mathcal{D},0}$ defined by (48) with $v = 0$, i.e.,

$$\forall K \in \mathcal{T},\ \forall x \in K :$$
$$\overline{\nabla}_{\mathcal{D},0} w(x) = \frac{1}{|K|} \mathbf{D}_K^{-1} \sum_{a\in\mathcal{A}_K} |a| \frac{\epsilon_{1,K} F_{1,a}(w) + \epsilon_{2,K} F_{2,a}(w)}{2} (X_K - \bar{x}_a).$$

4. Implementation and numerical behavior of the scheme $\mathcal{S}$.

4.1. About the choice of the various parameters. The practical construction of $\mathcal{S}$, following the recipe of section 2, requires us to make several choices of the points $M_{i,a}$ and of the parameters $\beta_{i,a}$ in particular. The theoretical study made in section 3 can give some insights on how to make proper choices for these quantities.

Choice of $\boldsymbol{\beta_{i,a}}$. As noticed in section 3.1, the continuity of the function defining the scheme is not ensured unless (28) is satisfied. In all probability, the computation and numerical stability of a solution to the scheme requires this continuity, and it therefore seems natural to take $\beta_{1,a} = \beta_{2,a}$ for all edges a.

Choice of the convex combination defining $\boldsymbol{M_{i,a}}$. The proof of Lemma 3.5 shows that the consistency of the numerical fluxes on regular functions is, in general, ensured only for edges across which $\mathbf{D}$ is continuous. This is not a surprise: the numerical fluxes $F_{K,a}$ are built to be conservative across all the edges of the mesh, but, if $\varphi \in C^2$, the fluxes $(-\mathbf{D}\nabla\varphi)\cdot\vec{n}$ on an edge corresponding to a discontinuity of $\mathbf{D}$ *are not* conservative in general ($\nabla\varphi$ has no jump across the edge and therefore cannot compensate for a jump of $\mathbf{D}$); there is thus no hope that $F_{i,a}(\varphi_{\mathcal{D}})$ is a proper approximation of $(-\mathbf{D}\nabla\varphi)\cdot\vec{n}$ for such edges. As it has already been noticed (see, e.g., [13]) and as we saw during the study of the scheme, the nonconsistency of the fluxes on some particular edges—here the ones around the discontinuities of $\mathbf{D}$—does not prevent us from proving the convergence of the scheme (but demands a special treatment).

Let us, however, look in more detail into this consistency issue. The technical reason is the following: if φ is regular, the values $\varphi_{T_i(a),a}$ computed by imposing the

conservativity of the fluxes are not order 2 approximations of $\varphi(X_{T_i(a),a})$. This conservativity is quite mandatory in finite volume methods, and it is moreover physical: whatever the discontinuities of $\mathbf{D}$, the fluxes $(-\mathbf{D}\nabla u)\cdot\vec{n}$ of the exact solution u to (1) are themselves always conservative. The choice of the values at $X_{T_i(a),a}$ can therefore hardly be changed, and in general these values will not be proper approximations for regular functions. But if we consider continuous and piecewise-regular functions φ, with jumps of derivatives aligned along the discontinuities of $\mathbf{D}$ and such that the fluxes $(-\mathbf{D}\nabla\varphi)\cdot\vec{n}$ are conservative across all the edges (the exact solution to (1) satisfies in particular these properties), then there is no longer any nonconservativity objection to the consistency of the numerical fluxes for φ. However, another objection arises: for piecewise-regular functions, the interpolations (17) used to construct the fluxes are not order 2 approximations of the function at $M_{i,a}$, and the fluxes are therefore once again not consistent.

What we notice here is in fact an intrinsic contradiction in the general construction made in section 2.2: the convex combinations (17) are justified for regular functions, whereas the choice of the edge values (20) assumes that all the fluxes are conservative (even if $\mathbf{D}$ is discontinuous). Both assumptions cannot hold in general. Though this does not prevent the scheme from converging, this "contradictory" construction probably leads to less precise results on coarse meshes.

There is, however, a way to make proper choices of parameters that eliminate this apparent contradiction. If $\mathbf{D}$ is piecewise-C^2, it is proved in [3] that the continuous functions ϕ that are C^2 on the same subdomains as $\mathbf{D}$ and such that the fluxes $(-\mathbf{D}\nabla\phi)\cdot\vec{n}$ are continuous (even across the discontinuities of $\mathbf{D}$) are dense in $H^1_0(\Omega)$; these functions therefore provide enough test functions to prove the convergence of the scheme and do not present the nonconservativity issue preventing their numerical fluxes from being proper approximations of their exact fluxes. Assume now that each convex combination (16) is written using only points $(M_{i,a}, X_{T_i(a)}, (X_{i,a}(j))_{i=1,J_{i,a}})$ that belong to the same subdomain on which $\mathbf{D}$ is C^2; all the functions ϕ from [3] being C^2 on the same subdomains as $\mathbf{D}$, this choice ensures that $\phi_{M_{i,a}}$ given by (17) is an order 2 approximation of $\phi(M_{i,a})$. As in the proof of Lemma 3.5, we then see, thanks to the conservativity of the fluxes $(-\mathbf{D}\nabla\phi)\cdot\vec{n}$, that $\phi_{T_i(a),a}$ computed by (20) is an order 2 approximation of $\phi(X_{T_i(a),a})$, and thus that the numerical flux $F_{i,a}(\phi_{\mathcal{D}})$ is a consistent approximation of the continuous flux (even in the presence of discontinuities of $\mathbf{D}$). This leads to the following principle, which should be satisfied whenever possible (i.e., when the mesh is not too coarse or $\mathbf{D}$ does not have too many discontinuities):

(50) $M_{i,a}$ and all the $(X_{T_i(a)}, (X_{i,a}(j))_{j=1,J_{i,a}})$ used to write the convex combination (16) belong to the same subdomain on which $\mathbf{D}$ is regular.

Besides the improvement in the quality of the scheme (see the next section), such a choice avoids the special handling, in the proof of Theorem 3.8, of the edges around the discontinuities of $\mathbf{D}$. As we noticed above, satisfying (50) can be impossible on too coarse meshes (e.g., when the scale of heterogeneity of $\mathbf{D}$ is the same as the scale of the mesh); however, even if (50) cannot be satisfied on some edges, the proof above shows that, as the mesh size tends to 0, the method still converges[2]—but possibly with a degraded order of convergence (see the numerical tests following). We do not know

[2]Indeed, as we explain above and as noticed in [22, 13], the nonconsistency of the fluxes on some edges—here the ones around the discontinuities of $\mathbf{D}$, and on which it might be impossible to satisfy (50)—does not necessarily prevent the scheme from converging as the mesh size tends to 0.

of a perfectly reliable method that ensures that the fluxes can be approximated in a consistent way on coarse meshes and for any heterogeneity of $\mathbf{D}$; nevertheless, in this case, a possible additional improvement can be to introduce the special interpolations of [4]. These interpolations allow us on some grids to construct edge approximations of order 2 of the solution despite the discontinuity of the tensor across the edges; using these approximate edge values in (17) as additional $u_{X_{i,a}(j)}$, one can more easily find order 2 approximations of the solution at $M_{i,a}$ even on coarse meshes. Such a study will be the subject of future works.

4.2. Numerical results. The nonlinear scheme $\mathcal{S}$ is solved by implementing the simple iterative algorithm described in [33]: we fix to $u = u^n$ the argument of $\gamma_{1,a}$ and $\gamma_{2,a}$ in (26) and solve for $u = u^{n+1}$ the corresponding linear scheme (27); the convergence criterion is achieved when the relative difference in the L^2 norm between two iterations u^n and u^{n+1} is less than 10^{-5}.

Remark 4.1. Obviously, the scheme being nonlinear (which is nearly a requirement for monotone methods; see the introduction), the computation of its solution is more costly than for linear methods. We will nevertheless see that, in many tests, the number of iterations in the algorithm stays relatively low; it should also be noticed that the linear system solved at each iteration comes from an M-matrix that is roughly as well conditioned as the matrices of usual linear schemes and is also well adapted to multigrid methods. Finally, we notice that realistic models (see section 4.2.3) are nonlinear and that, in practice, the "additional" nonlinearity introduced by our nonlinear method increases very little the cost of solving the scheme.

In [31] and [33], the scheme $\mathcal{S}$ has been tested on the benchmark of [25]. The analytical tests (test 1 and 5) show second order accurate results, as for the classical linear schemes; the algorithm described above to solve $\mathcal{S}$ converges in less than 10 iterations. On the stiffest tests (tests 8 and 9), about 200 iterations are needed to solve $\mathcal{S}$, but the solutions obtained with all the other linear schemes show large oscillations, and are therefore much less acceptable than the one given by $\mathcal{S}$.

Other numerical results on $\mathcal{S}$ are provided in previous references [30, 33] in the case of regular tensors $\mathbf{D}$. In the following numerical results, we therefore consider situations in which $\mathbf{D}$ is strongly discontinuous and anisotropic, which was not truly the case in most of the preceding tests. We concentrate mainly on two specific questions: (i) comparison of $\mathcal{S}$ (having the LMP structure) with the simpler linear version, denoted by $\mathcal{L}$ and lacking the LMP structure, obtained by choosing $\gamma_{1,a}(u) = \gamma_{2,a}(u) = \frac{1}{2}$ in (26); and (ii) influence of different possible choices for the points $M_{i,a}$.

We consider six grids, described in [25] (see Mesh1) that contain from 56 triangular cells (the first) to 57000 triangular cells (the sixth). As explained above, the coercivity assumption (29) cannot, in general, be theoretically verified; however, some numerical quantities can be computed in order to check whether this assumption seems to hold or not. Rather than trying to obtain a numerical value for the ζ in (29), we prefer to consider

$$\nu_h = \frac{\sum_{a\in\mathcal{A}} |a| \min\left[\epsilon_{1,K} F_{1,a}(u)(u_L - u_K); \epsilon_{2,K} F_{2,a}(u)(u_L - u_K)\right]}{||u||_{\mathcal{D}}^2},$$

where u is the solution to $\mathcal{S}$; ν_h is easy to compute and gives a good indication of the coercivity of the method in the region of interest: if ν_h stays positive and far from 0 as the mesh size h tends to 0, then one can consider that the method is coercive at least on a neighborhood of the solution.

Some notation used to present the numerical results is given in Table 1.

TABLE 1
Notation.

h	Size of the discretization (i.e., shortcut for size($\mathcal{D}$))
L^2 error	L^2 error of the computed solution with respect to the analytical solution
ratiol2	Order of convergence, in L^2 norm, of the method
nit	Number of iterations needed to compute the approximate solution of $\mathcal{S}$
nmp	Number of iterations, in the solving of $\mathcal{S}$, before the iterative solution u^n satisfies the maximum principle (i.e., takes its values between the minimum and maximum values of the boundary data)
ν_h	Coercivity parameter as a function of $h = \text{size}(\mathcal{D})$

4.2.1. Stationary analytical solution. We consider the following elliptic problem:

$$\left\{ \begin{array}{ll} \operatorname{div}(\mathbf{D}\nabla u) = \operatorname{div}(\mathbf{D}\nabla u_{\text{ana}}) & \text{in } \Omega =]0,1[\times]0,1[\,, \\ u = u_{\text{ana}} & \text{on } \partial\Omega \end{array} \right. \tag{51}$$

with

$$\mathbf{D}(x,y) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ if } x \leq 0.5\,, \quad \mathbf{D}(x,y) = \begin{pmatrix} 100 & 0 \\ 0 & 0.01 \end{pmatrix} \text{ if } x > 0.5, \tag{52}$$

and the analytical solution

$$u_{\text{ana}}(x,y) = \cos(\pi x)\sin(\pi y) \text{ if } x \leq 0.5\,, \quad u_{\text{ana}}(x,y) = 0.01\cos(\pi x)\sin(\pi y) \text{ if } x > 0.5.$$

This analytical solution and its fluxes are continuous at the interface $x = 0.5$, so that $\operatorname{div}(\mathbf{D}\nabla u_{\text{ana}}) \in L^2(\Omega)$.

In the first two numerical tests, we look at the linear scheme $\mathcal{L}$. Table 2 presents the results for when the principle (50) is not taken into account, and Table 3 shows the results for when this principle is respected. Since the exact solution is not C^2 in the whole domain, as expected the convergence is much better if (50) holds: the order is in fact twice as large.

TABLE 2
Linear scheme $\mathcal{L}$ for (51)*, principle* (50) *not respected.*

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
L^2 error	8.4×10^{-2}	3.4×10^{-2}	1.5×10^{-2}	7.0×10^{-3}	3.3×10^{-3}	1.6×10^{-3}
ratiol2		1.27	1.16	1.11	1.09	1.03

TABLE 3
Linear scheme $\mathcal{L}$ for (51)*, principle* (50) *respected.*

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
L^2 error	3.3×10^{-2}	8.9×10^{-3}	2.2×10^{-3}	5.6×10^{-4}	1.4×10^{-4}	3.4×10^{-5}
ratiol2		1.89	1.98	2.00	2.02	2.04

We then present in Tables 4 and 5 the results for the nonlinear scheme $\mathcal{S}$, choosing $\beta_{1,a} = \beta_{2,a} = \min(\frac{\alpha_{1,a}}{\delta_{1,a}}, \frac{\alpha_{2,a}}{\delta_{2,a}})$. The difference between the two tests still consists of respecting, or not, respecting the principle (50). Once again, it is clear that this principle genuinely improves the quality of the scheme, doubling its order of convergence; moreover, it is interesting to notice that this principle also largely accelerates the

TABLE 4
Nonlinear scheme $\mathcal{S}$ for (51)*, principle* (50) *not respected.*

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
L^2 error	3.4×10^{-2}	8.4×10^{-3}	2.3×10^{-3}	7.7×10^{-4}	3.4×10^{-4}	1.7×10^{-4}
ratiol2		2.00	1.88	1.57	1.17	1.01
nit	88	124	99	76	52	28
ν_h	1.05	0.93	0.98	1.03	1.06	1.08

TABLE 5
Nonlinear scheme $\mathcal{S}$ for (51)*, principle* (50) *respected.*

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
L^2 error	3.4×10^{-2}	8.6×10^{-3}	2.2×10^{-3}	5.5×10^{-4}	1.4×10^{-4}	3.4×10^{-5}
ratiol2		1.96	1.96	2.00	2.00	2.03
nit	110	43	7	4	3	2
ν_h	0.94	1.13	1.11	1.10	1.10	1.10

TABLE 6
Nonlinear scheme $\mathcal{S}$ for (53)*.*

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
nit	117	124	92	73	59
nmp	2	3	2	2	1
ν_h	0.70	1.60	2.44	3.03	3.22

convergence of the fixed point algorithm. The choice of (50) therefore not only gives a more precise approximate solution, but also an easier one to compute in practice.

These tables also show that, in all the tests, the quantity ν_h stays positive and does not tend toward zero; this numerically suggests that assumption (29) holds quite well in practice around the solution of the problem, and thus that we are in the theoretical framework of convergence developed in section 3.

Remark 4.2. We also tested, for the nonlinear scheme, the choices of $\beta_{1,a} = \frac{\alpha_{1,a}}{\delta_{1,a}}$ and $\beta_{2,a} = \frac{\alpha_{2,a}}{\delta_{2,a}}$; in general, $\beta_{1,a}$ and $\beta_{2,a}$ are then different, and we observe that the fixed point algorithm does not converge. This confirms the requirement $\beta_{1,a} = \beta_{2,a}$ when implementing $\mathcal{S}$.

4.2.2. Stationary nonanalytical solution. In order to evaluate the respect of the discrete maximum principle, we now consider the following problem:

$$\text{(53)} \qquad \begin{cases} \operatorname{div}(\mathbf{D}\nabla u) = 0 & \text{in } \Omega =]0,1[\times]0,1[\,, \\ u = x & \text{on } \partial\Omega, \end{cases}$$

where $\mathbf{D}$ is as before (see (52)).

Table 6 and Figure 4 show the results of $\mathcal{S}$ when applying the principle (50). We notice that nit is much larger than with the previous analytical solution. A possible explanation for this is the following: as we show below, the concentrations obtained with the linear schemes for (53) are very oscillating (more than in the previous test); these oscillations indicate that the considered problem is stiffer than the preceding one, and it could explain why the nonlinear method requires more iterations in the fixed point algorithm to achieve the convergence criterion.

In Table 7 and Figure 5, we show the results on the same problem (53) with the VFSYM scheme of [29] (a symmetric cell centered finite volume scheme on simplexes,

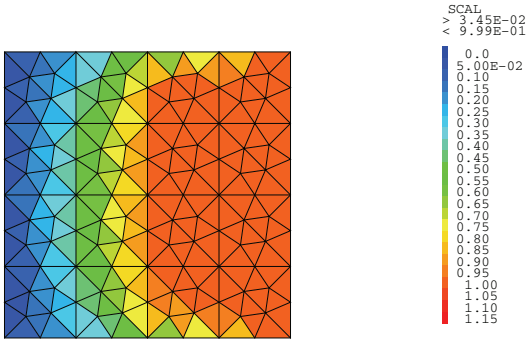

FIG. 4. *Concentration given by* $\mathcal{S}$ *for* (53) *on a grid made of* 224 *cells (maximum value* 0.99*, minimum value* 0.03*).*

TABLE 7

VFSYM scheme for (53)*: percentage of overshoots and maximum values of the approximate solutions.*

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
Overshoots	32%	19%	10%	4%	1.0%
$u_{\max}$	1.09	1.09	1.07	1.06	1.03

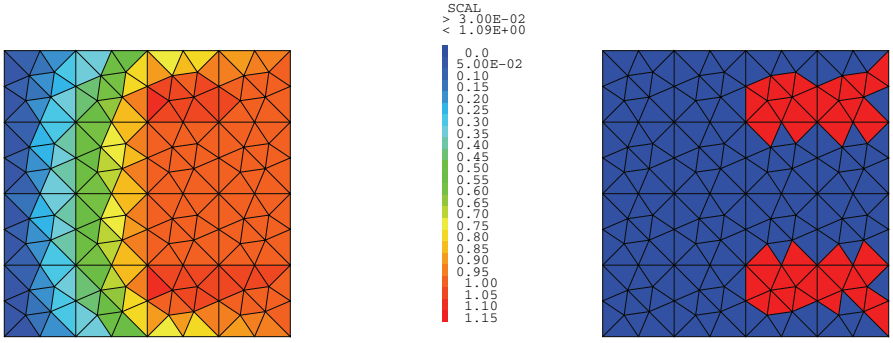

FIG. 5. *Concentration and position of the overshoots (in red) for the VFSYM scheme on* (53) *on a grid made of* 224 *cells (maximum value* 1.09*, minimum value* 0.03*). See online version for color.*

parallelograms, and parallelepipeds); this table and figure also present the percentage of values higher than 1 (upper bound of the boundary data for (53)) and the maximum values of the approximate solutions.

Finally, Table 8 and Figure 6 present the behavior of the linear scheme $\mathcal{L}$ on (53).

It is interesting to observe that the two linear schemes (VFSYM and $\mathcal{L}$) present oscillations on all grids; these oscillations can be quite large and numerous unless the grid is thin. On the other hand, as expected, no such oscillations appear with the nonlinear scheme $\mathcal{S}$; moreover, fewer than three iterations are required in the fixed point

TABLE 8

Linear scheme $\mathcal{L}$ for (53): percentage of overshoots and maximum values of the approximate solutions.

h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
Overshoots	35%	21%	11%	5%	0.8%
$u_{\max}$	1.11	1.13	1.13	1.09	1.03

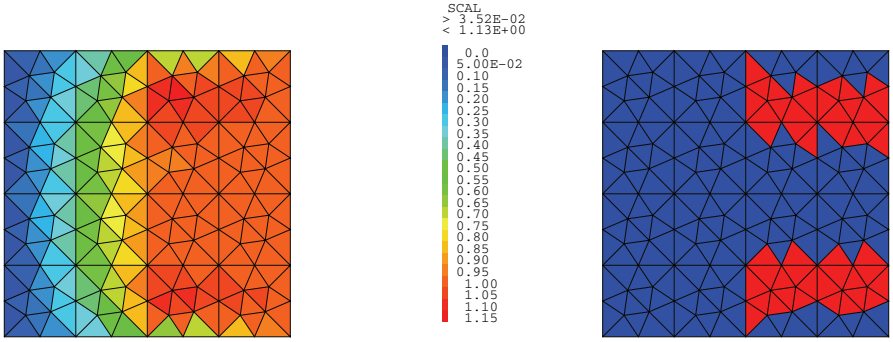

FIG. 6. *Concentration and position of the overshoots (in red) for the linear scheme $\mathcal{L}$ on (53) on a grid made of 224 cells (maximum value 1.13, minimum value 0.0352). See online version for color.*

algorithm to obtain the minimum and maximum principles: hence, not only does the exact solution to this scheme satisfy the discrete maximum and minimum principles, but the practical approximations we manage to compute also almost immediately exhibit no overshoots or undershoots.

4.2.3. The ANDRA COUPLEX 1 test case. We now consider $\mathcal{S}$ on the ANDRA COUPLEX 1 test case [7]. More specifically, we study the transport of the Iodine 129 that escapes from a repository cave into the water. The concentration C satisfies the following convection-diffusion equation:

$$\omega(\partial_t C + \lambda C) - \operatorname{div}(\mathbf{Diff}\nabla C + \vec{\mathrm{vel}}\, C) = f \quad \text{on } \Omega\times]0,T[, \tag{54}$$

where **Diff** is the diffusion-dispersion tensor, ω is the effective porosity, and $\lambda = \frac{\log(2)}{T_{\text{half}}}$ (T_{half} being the half-life of Iodine 129). The Darcy velocity $\vec{\mathrm{vel}} = \mathbf{K}\nabla h$ satisfies the flow equation $\operatorname{div}(\vec{\mathrm{vel}}) = 0$, where $\mathbf{K}$ is the permeability tensor and h is the dynamic load. All the coefficients are given in [7].

We consider a grid of the domain Ω made of about 3600 triangular cells. We compute the Darcy velocity by solving the flow equation using the finite volume scheme described in [1].

We then consider two discretizations of (54). Both are based on a time-implicit scheme and use an upwind finite volume discretization of the convection term; they differ only in the discretization of the diffusion-dispersion tensor: the first uses the multipoint flux approximation (MPFA) of [1], and the second applies the nonlinear method $\mathcal{S}$. In both cases, we use 200-year time steps from 1000 to 2000 years, 500-year time steps from 2000 to 10110 years, 1250-year time steps from 10110 to 50110

years, 5000-year time steps from 50110 to 2×10^5 years, and 10000-year time steps from 2×10^5 to 10^6 years (the total is about 165 time steps).

Figures 7–10 show the concentration given by the two methods: obviously, $\mathcal{S}$ suppresses all the spurious oscillations present in the MPFA method; moreover, in the zone where the MPFA scheme does not oscillate, the results obtained with the two methods are very similar. It is also interesting to notice that the cost of solving the nonlinear scheme $\mathcal{S}$ on this test case is quite low: at most four iterations for each time step. Some very tiny oscillations (inferior to 1E-15) can be seen in the solution given by $\mathcal{S}$; they are a consequence of the limited machine precision.

These results clearly show that the nonlinear method we constructed and studied in (1) also behaves well, and with a limited cost, on more realistic and complex models.

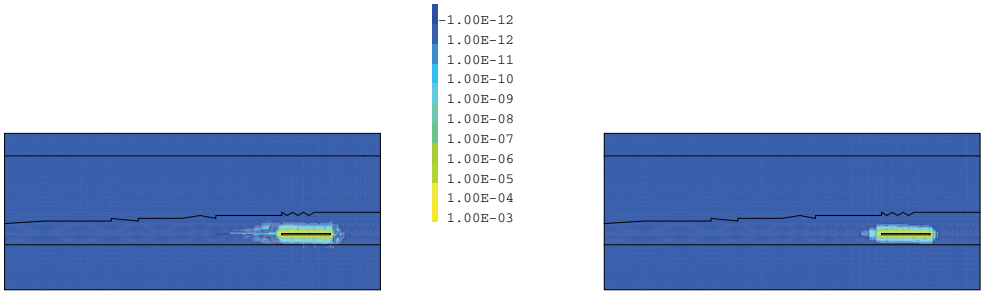

FIG. 7. *Contour levels of iodine concentration at* 10110 *years. Left: MPFA scheme (maximum value* 1.14×10^{-4}*, minimum value* -3.86×10^{-6}*). Right:* $\mathcal{S}$ *(maximum value* 1.08×10^{-4}*, minimum value* -5.67×10^{-16}*).*

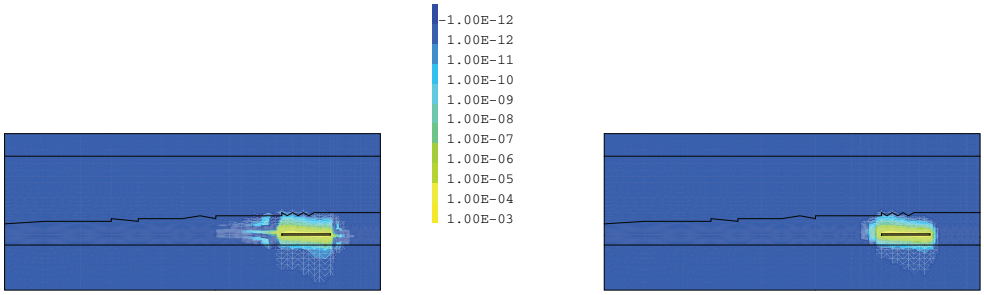

FIG. 8. *Contour levels of iodine concentration at* 50110 *years. Left: MPFA scheme (maximum value* 1.69×10^{-4}*, minimum value* -4.37×10^{-6}*). Right:* $\mathcal{S}$ *(maximum value* 1.62×10^{-4}*, minimum value* -3.65×10^{-17}*).*

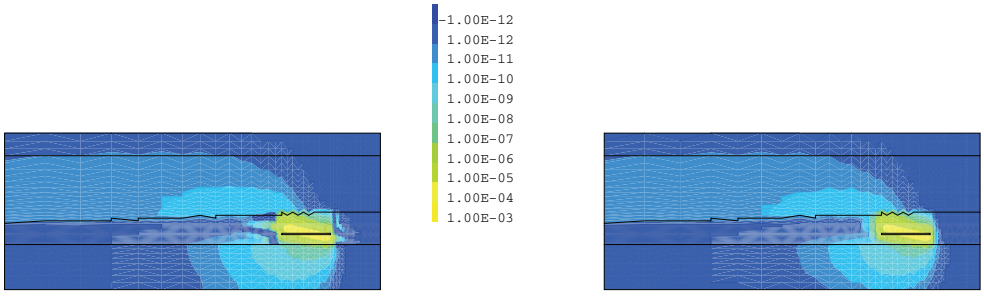

FIG. 9. *Contour levels of iodine concentration at* 200000 *years. Left: MPFA scheme (maximum value* 1.03×10^{-4}*, minimum value* -2.23×10^{-6}*). Right:* $\mathcal{S}$ *(maximum value* 1.12×10^{-4}*, minimum value* -2.06×10^{-17}*).*

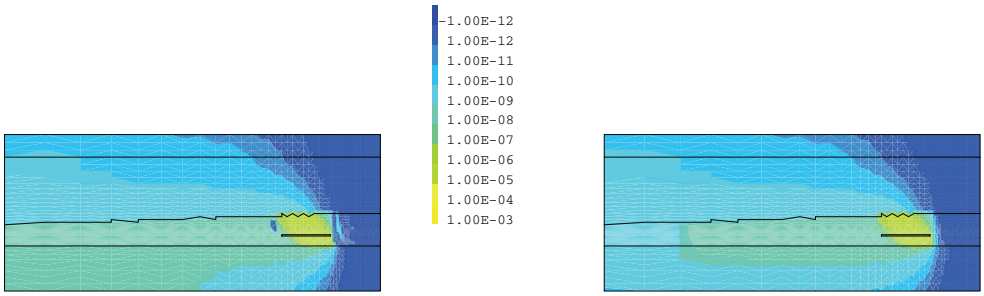

FIG. 10. *Contour levels of iodine concentration at* 1000000 *years. Left: MPFA scheme (maximum value* 3.81×10^{-5}*, minimum value* -1.86×10^{-7}*). Right:* $\mathcal{S}$ *(maximum value* 4.23×10^{-5}*, minimum value* 1.50×10^{-29}*).*

5. Conclusion. We presented a numerical method for diffusion equations that respects the minimum and maximum principles and is nonoscillating. This method is based on a nonlinear combination of linear fluxes, and it can be constructed in two or three dimensions on very generic grids and in the presence of strong anisotropy and heterogeneity. We made a theoretical study of the method, proving in particular its convergence under a coercivity assumption; to the best of our knowledge, very few monotone schemes have been proved to be convergent on generic grids. This study allowed us in particular to understand how to properly choose the parameters in the presence of discontinuous tensors; the case of discontinuity in the diffusion tensor is obviously of utmost importance in practical situations but is scarcely considered in the literature on monotone schemes. The numerical results confirmed the theoretical predictions and the good behavior of the method. Despite the nonlinearity of the scheme, the computation of the approximate solution is in many cases not too difficult. Finally, the tests on the more challenging COUPLEX 1 benchmark also showed that our method is promising for more complex models than pure linear diffusion equations.

REFERENCES

[1] I. Aavatsmark, T. Barkve, O. Bøe, and T. Mannseth, *Discretization on unstructured grids for inhomogeneous, anisotropic media. Part* I: *Derivation of the methods*, SIAM J. Sci. Comput., 19 (1998), pp. 1700–1716.

[2] I. Aavatsmark, T. Barkve, O. Bøe, and T. Mannseth, *Discretization on unstructured grids for inhomogeneous, anisotropic media. Part* II: *Discussion and numerical results*, SIAM J. Sci. Comput., 19 (1998), pp. 1717–1736.

[3] L. Agelas, D. Di Pietro, and J. Droniou, *The G method for heterogeneous anisotropic diffusion on general meshes*, M2AN Math. Model. Numer. Anal., 44 (2010), pp. 597–625.

[4] L. Agelas, R. Eymard, and R. Herbin, *A nine-point finite volume scheme for the simulation of diffusion in heterogeneous media*, C. R. Acad. Sci. Paris Ser. I, 347 (2009), pp. 673–676.

[5] L. Beirão da Veiga and G. Manzini, *A higher-order formulation of the mimetic finite difference method*, SIAM J. Sci. Comput., 31 (2008), pp. 732–760.

[6] E. Bertolazzi and G. Manzini, *A second-order maximum principle preserving finite volume method for steady convection-diffusion problems*, SIAM J. Numer. Anal., 43 (2005), pp. 2172–2199.

[7] A. Bourgeat, M. Kern, S. Schumacher, and J. Talandier, *The COUPLEX Test Cases: Nuclear Waste Disposal Simulation,* http://www.gdrmomas.org/Ex_qualif/Couplex/Documents/couplexall.pdf (2002).

[8] F. Boyer and F. Hubert, *Finite volume method for* 2*D linear and nonlinear elliptic problems with discontinuities*, SIAM J. Numer. Anal., 46 (2008), pp. 3032–3070.

[9] F. Brezzi, K. Lipnikov, and V. Simoncini, *A family of mimetic finite difference methods on polygonal and polyhedral meshes*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1533–1551.

[10] F. Brezzi, K. Lipnikov, and M. Shashkov, *Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes*, SIAM J. Numer. Anal., 43 (2005), pp. 1872–1896.

[11] C. Buet and S. Cordier, *On the nonexistence of monotone linear schema for some linear parabolic equations*, C. R. Acad. Sci. Paris Ser. I, 340 (2005), pp. 399–404.

[12] E. Burman and A. Ern, *Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes*, C. R. Acad. Sci. Paris Ser. I, 338 (2004), pp. 641–646.

[13] R. Cautres, R. Herbin, and F. Hubert, *The Lions domain decomposition algorithm on non-matching cell-centered finite volume meshes*, IMA J. Numer. Anal., 24 (2004), pp. 465–490.

[14] C. Chainais-Hillairet, J.-G. Liu, and Y.-J. Peng, *Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 319–338.

[15] Y. Coudière, J. P. Vila, and P. Villedieu, *Convergence rate of a finite volume scheme for a two dimensional convection diffusion problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 493–516.

[16] K. Deimling, *Nonlinear Functional Analysis*, Springer, New York, 1985.

[17] D. A. Di Pietro and A. Ern, *Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible Navier–Stokes equations*, Math. Comp., 79 (2010), pp. 1303–1330.

[18] K. Domelevo and P. Omnes, *A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 1203–1249.

[19] J. Droniou and R. Eymard, *A mixed finite volume scheme for anisotropic diffusion problems on any grid*, Numer. Math., 105 (2006), pp. 35–71.

[20] J. Droniou, R. Eymard, T. Gallouët, and R. Herbin, *A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods*, Math. Models Methods Appl. Sci., 20 (2010), pp. 265–295.

[21] J. Droniou and R. Eymard, *Study of the mixed finite volume method for Stokes and Navier-Stokes equations*, Numer. Methods Partial Differential Equations, 25 (2009), pp. 137–171.

[22] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis VII, Techniques of Scientific Computing, Part III, P. G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 2000, pp. 713–1020.

[23] R. Eymard, T. Gallouët, and R. Herbin, *Discretisation of heterogeneous and anisotropic diffusion problems on general non-conforming meshes. SUSHI: A scheme using stabilisation and hybrid interfaces*, IMA J. Numer. Anal., 30 (2010), pp. 1009–1043.

[24] R. Eymard, T. Gallouët, and R. Herbin, *Cell centered discretization of nonlinear elliptic problems on general multidimensional polyhedral cells*, J. Numer. Math., 17 (2009), pp. 173–193.

[25] R. Herbin and F. Hubert, *Benchmark on discretization schemes for anisotropic diffusion problems on general grids*, in Finite Volumes for Complex Applications V, R. Eymard and J.-M. Hérard, eds., ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, 2008, pp. 659–692.

[26] I. Kapyrin, *A family of monotone methods for the numerical of three-dimensional diffusion problems on unstructured tetrahedral meshes*, Dokl. Math., 76 (2007), pp. 734–738.

[27] E. Keilegavlen, J. M. Nordbotten, and I. Aavatsmark, *Sufficient criteria are necessary for monotone control volume methods*, Appl. Math. Lett., 22 (2009), pp. 1178–1180.

[28] D. S. Kershaw, *Differencing of the diffusion equation in Lagrangian hydrodynamic codes*, J. Comput. Phys., 39 (1981), pp. 375–395.

[29] C. Le Potier, *Schéma volumes finis pour des opérateurs de diffusion fortement anisotropes sur des maillages non structurés*, C. R. Acad. Sci. Paris Ser. I, 340 (2005), pp. 921–926.

[30] C. Le Potier, *Schéma volumes finis monotone pour des opérateurs de diffusion fortement anisotropes sur des maillages de triangles non structurés*, C. R. Acad. Sci. Paris Ser. I, 341 (2005), pp. 787–792.

[31] C. Le Potier, *Numerical results with two cell-centered finite volume schemes for heterogeneous anisotropic diffusion operators*, in Finite Volumes for Complex Applications V, R. Eymard and J.-M. Hérard, eds., ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, 2008, pp. 825–842.

[32] C. Le Potier, *Un schéma linéaire vérifiant le principe du maximum pour des opérateurs de diffusion très anisotropes sur des maillages déformés*, C. R. Acad. Sci. Paris Ser. I, 347 (2009), pp. 105–110.

[33] C. Le Potier, *A nonlinear finite volume scheme satisfying maximum and minimum principles for diffusion operators*, Int. J. Finite Vol., 6 (2009), no. 2.

[34] K. Lipnikov, M. Shashkov, D. Svyatskiy, and Yu. Vassilevski, *Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes*, J. Comput. Phys., 227 (2007), pp. 492–512.

[35] K. Lipnikov, D. Svyatskiy, and Yu. Vassilevski, *Interpolation-free monotone finite volume method for diffusion equations on polygonal meshes*, J. Comput. Phys., 228 (2009), pp. 703–716.

[36] J. M. Nordbotten, I. Aavastsmark, and G. T. Eigestad, *Monotonicity of control volume methods*, Numer. Math., 106 (2007), pp. 255–288.

[37] B. Rivière, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation*, Frontiers Appl. Math. 35, SIAM, Philadelphia, 2008.

[38] G. Yuan and Z. Sheng, *Monotone finite volume schemes for diffusion equations on polygonal meshes*, J. Comput. Phys., 227 (2008), pp. 6288–6312.